KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

A fresh look at patent citations

Dissertation presented to obtain the degree of Doctor in Business Economics

by

Jurriën Bakker

Number: 590 2017

Daar de proefschriften in de reeks van de Faculteit Economie en Bedrijfswetenschappen het persoonlijk werk zijn van hun auteurs, zijn alleen deze laatsten daarvoor verantwoordelijk.

For Diana, for without you this would have not been possible.

Table of contents

List of figures

List of tables

Doctoral committee

Promotor

Prof. Dr. Bart Van Looy (KU Leuven)

Doctoral committee members

Prof. Dr. Dirk Czarnitzki
(KU Leuven)

Prof. Dr. Otto Toivanen
(KU Leuven)

Prof. Dr. Koen Frenken
(Utrecht University)

Prof. Dr. Dietmar Harhoff
(Max Planck Institute for Innovation and Competition)

Acknowledgements

Making this PHD was a long journey, which passed through major hurdles, but fortunately also involved great successes. This would have not been possible without all of those who supported me. In this section, I would like to extend my gratitude to all of you.

First and foremost, I would like to thank my wife Diana Ferreira, for without her nothing like this would have been possible. Besides her invaluable moral support, she greatly helped me by applying her legal and business acumen to the world of intellectual property. I would also like to thank my parents Joke and Jaap Bakker, my sister Nina Bakker, her partner Fabian de Jong, and my parents-in law Fernando Moutinho and Fernanda Silva for providing me with guidance and support during this journey.

Next, I would like to extend my gratitude to the advisors of my dissertation. My promotor Bart Van Looy helped me kickstart my thesis by pointing to the problems that are frequently ignored in the field of innovation measurement. My committee members Dirk Czarnitzki and Otto Toivanen helped me with their comments, while my external committee members Koen Frenken and Dietmar Harhoff managed to provide me with just the right questions which helped me rethink my thesis. I would also like to thank my collaborators: Dennis Verhoeven and Lin Zhang, who helped me write and publish the first chapter of this dissertation.

My stay at the KU Leuven was made a lot more pleasant and productive by the people that helped me. For instance, Manuel Gigena, who helped me by providing code and an understanding of great coffee. There is also Adrián Kovács, who was a friendly and helpful Dutch presence throughout my PHD. I am also gratefull to Tom Magerman, Federico de Michiel, Paul-Emmanuel Anckaert, Daniela Silvestri, Jan-Bart Vervenne and Antonio Della Malva for providing me with help for my projects.

Finally, I would also like to thank Cindy Lopes Bento, Thomas Schaper, Maikel Pellens, Linde Colen, Wytse Joosten, Maarten Rabijns, Marcelina Grabowska, Michela Bergamini, Kieran Dobson, Nazareno Braito, Hanne Peeters, Peter Smith, Sebastiaan Wijsman and many others for providing me with a warm and welcome atmosphere at the department of MSI of the Economic faculty of the KU Leuven.

Glossary

Applicant	Person or company applying for a patent
Censored sample	Sample that excludes information for some observations
EPO	European Patent Office
Highly cited patent	Patent that is cited exceptionally often
Home bias	When an indicator favors entities residing in the territory where its data stems from
Inventor	Person listed as having invented the technology that is patented
IPC	International Patent Classification
Patent citation	Reference in one patent to another patent
------Backward	Reference in the focal patent to another patent
------Forward	Reference to the focal patent by another patent
Patent family	Set of related patents
------DOCDB	Patent family consisting of patents with equal technical content
------INPADOC	Patent family that attempts to group patents relating to the same invention
Patent license	Legal agreement that allows another party to use patented technology
Patent opposition	Process initiated at a patent office to prevent a patent from being issued
Patent portfolio	Set of patents owned by the same entity
Patent renewal	Payment of maintenance fees to keep a patent in force
Patent transfer	Sale of patent from one entity to antoher
Patent value	Value attributed to a patent
-----Inventive value	Value related to the technological progress described by the patent
-----Market value	Value of patent for entities that do not own it
-----Private value	Value of a patent to its owner
-----Productive value	Value derived from a patent by producing and selling products protected by it
----Social value	Value of a patent (and possibly its technology) for society
PCT	(Patent filed through the) Paris Cooperation Treaty
Tobin's Q	Measure of success for firms
USPTO	United States Patent and Trademark Office
WIPO	World Intellectual Property Organization

Populaire samenvatting

“Als innovatie nauwkeurig gemeten, geclassificeerd en voorspeld zou kunnen worden; dan zou het uiteindelijk niet innovatief zijn ”

Innovatie is de drijvende kracht voor onze kenniseconomie. Het is daarom belangrijk om te begrijpen hoe innovatie werkt en waar het vandaan komt. Ook zouden we graag willen weten hoe bedrijven, maar ook landen er op dit moment voorstaan ten aanzien van hun innovatieve capaciteiten. Om deze vragen te beantwoorden is het nodig om innovatie goed te kunnen meten.

Normaal gesproken tellen we hiervoor octrooien, ook wel patenten genoemd. Maar er zit een enorm verschil in hun waarde, wat een simpele optelsom een onbetrouwbare indicatie van innovatie maakt. Daarom is het belangrijk om deze waarde te bepalen. De meest gebruikte methode hiervoor is het tellen van het aantal keer dat het octrooi geciteerd wordt door andere octrooien, ook wel octrooi citaties genoemd. Dit doctoraat is erop gericht om deze methode, onder de loep te nemen en, waar nodig, te verbeteren. Daarom zullen drie van de grootste problemen met de indicator worden besproken.

Het eerste probleem is dat er veel verschillende manieren zijn om te bepalen welke octrooien welke ander octrooien citeren. Dit komt vooral doordat onderzoekers verschillende methodes en databronnen gebruiken. De resultaten van dit doctoraat latent zien dat de waardebepaling drastisch kan verschillen afhankelijk van de opzet die de onderzoeker gebruikt. Daarnaast wordt er in dit doctoraat ook bepaald welke methode leidt tot de beste waardebepaling en dat dit blijkt te verschillen voor octrooien uit verschillende landen.

In de meeste gevallen wordt de waarde van een patent bepaald door simpelweg het aantal keer dat het is geciteerd bij elkaar op te tellen. Dit aantal citaties verschilt enorm tussen de octrooien, daarom wordt in het doctoraat de vraag gesteld of dit ook zulke grote verschillen in de waarde van de octrooien kenmerkt. Dit blijkt niet zo te zijn. Het is daarom de aanbeveling om de waarde van een octrooi te laten bepalen door de logaritme van de som van het aantal citaties, omdat deze functionele vorm de waarde verdeling beter benadert.

Tenslotte kunnen de ontvangen citaties nog worden gecategoriseerd afhankelijk van de eigenaar van de citerende octrooien. In dit geval wordt bekeken of de eigenaar hetzelfde is als die van het geciteerde octrooi, een ‘zelf-citatie’, of niet. Het blijkt vervolgens dat zelf-citaties een indicatie zijn van de waarde die de eigenaar hecht aan het exclusieve monopoly dat wordt verkregen door het octrooi. Niet-zelf-citaties geven meer een idee van de waarde van het octrooi op de markt voor intellectueel eigendom, omdat ze een indicatie zijn van mogelijke doelwitten die kunnen worden aangeklaagd voor inbreuk op het patent. Door deze typen citaties apart te bestuderen kunnen we een beter beeld krijgen van de strategieën van octrooi eigenaren.

De resultaten beschreven in dit doctoraat zouden, als ze worden overgenomen, kunnen leiden tot een verbetering van het meten van de waarde van octrooien. Dit is niet alleen nuttig voor eigenaars van intellectueel eigendom maar ook voor een beter begrip van octrooien en de innovatie waarop ze gebaseerd zijn.

Popular abstract

"If innovation could be accurately measured, classified and predicted; then it wouldn't be innovative at all"

Innovation is the driving force for our knowledge economy. It is therefore important to understand the inner workings of innovation as well as how it is generated. Additionally, it is important to know how companies and countries fare with respect to their innovative capacities. To measure this, as a rule, we simply count the number of patents filed by the relevant entity. But, there is a massive difference in the value of these patents, which makes this count an unreliable innovation indicator. Therefore, it is important to determine patent value. The most used indicators are obtained by counting the number of times the patent is referenced by other patents. This doctorate focusses on observing and, where necessary, improving this measure of determining patent value. Therefore, three important problems with the indicator will be discussed.

The first problem concerns the numerous ways that are used to determine which patents cite which other patents. Different researchers use different methods, which is mainly due to the various data sources and methods available. The results presented in this doctorate show that the valuation of patents can differ greatly depending on the methodological choices made by the researcher. Additionally, this doctorate determines which methods provide the best value estimates, which are different when patents from different countries are used.

Generally, patent value is determined by simply counting the number of times a patent is cited. The scores derived from this exercise differ greatly between patents. Therefore, this doctorate poses the question whether this is indicative of the large differences in value between patents. This appears not to be the case, leading to the recommendation of valuing patents using the logarithm of the count of their received citations, as this functional form better approaches the distribution of value within patents.

Finally, there is there is the notion that citations differ depending on the owner of the citing patent. Here, we determine if the owner is the same as the owner of the cited patent, in which case it is a self-citation. It is found that self-citations are an indication of the value that owners give to the temporary monopoly that the patent provides. Non-self-citations on the other hand are more an indication of the value of the patent in the market for intellectual property because they indicate potential targets that can be sued for infringing on the patent. By studying these types of citations separately, it is possible to better understand the strategies of patent owners.

The results described in this doctorate could, when they are adopted, lead to improvements in the measurement of patent value. This is not only useful for owners of this intellectual property but also for a better understanding of patents and the innovation on which they are based.

Samenvatting

Het aantal keren dat aan een octrooi wordt gerefereerd in andere octrooien, wordt over het algemeen gezien als een goede indicatie van zijn waarde en zijn innovatieve bijdrage. Maar er zijn drie cruciale problemen met deze indicator, beter bekend als octrooi citaties, die zijn gebruik hinderen en de validiteit van de studies, waarin hij wordt gebruikt, bedreigen. Deze doctors these zal deze problemen bespreken en oplossingen voorstellen.

Het eerste probleem betreft de diverse methoden die gebruikt worden om octrooi citaties te berekenen. De resultaten in deze scriptie laten zien dat deze methoden substantiële verschillende indicatoren produceren en dat over het algemeen de indicatoren die gebaseerd zijn op patent families, d.w.z. groepen van gerelateerde patenten, aan te raden zijn. Het tweede resultaat in de scriptie is dat de functionele vorm van de relatie tussen octrooi citaties en octrooi waarde beter beschreven kan worden met een log lineaire vorm dan met de gebruikelijke lineaire vorm. Tenslotte, de relatie tussen octrooi citaties en octrooi waarde is vooral empirisch gevonden of gebaseerd op een analogie met academische citaties. In deze these wordt een nieuwe theorie beschreven die de waarde van octrooi relateert aan de juridische context van octrooi citaties.

De resultaten van de thesis zouden, als ze worden overgenomen door onderzoekers die werken met octrooien, nuttig moeten zijn in het verbeteren van het meten van de relevantie van patenten. Hierdoor zou het begrip van innovatie en intellectueel eigendom moeten toenemen.

Abstract

The number of times a patent is referenced by other patents is generally seen as a good indication of its value and its innovative contribution. However, there are three key issues with the practical application of this indicator, better known as patent citations, which obfuscate its use and thus threaten the validity of studies in which it is applied. This doctoral thesis will discuss and remedy these three problems.

The first issue concerns the various methods that are in use to calculate patent citations. The results presented in this thesis indicate that these methods produce substantially different indicators and that, generally, indicators based on patent families, i.e. groups of related patents, present themselves as the preferred option. Next, it is found that patent citations and patent value are related through a log-linear relation, as opposed to the often-used linear relation. Starting from the legal and procedural role of citations, this thesis posits that patent citations relate to economic value albeit in two, distinctive, ways. These assertions are validated and empirically confirmed in this thesis.

Overall, the results of this thesis are relevant to improve the measurement and understanding of patent citations. By doing so, this should further the understanding of innovation and intellectual property.

General introduction

Why observe patent statistics?

Innovation, the concept of technological change, is the primary driver of economic growth. Within the field of economics, innovation has been recognized as an important economic phenomenon due to the work of Schumpeter (1942). Many economic efforts recognized that innovation is the most transformative force of historical societal change (Mokyr,1992), as well as the engine of economic growth (Aghion and Howitt,1990). In recent years, there has been an even greater emphasis on the study of innovation because many see in it the key to sustained economic growth and enhanced company performance.

In order to effectively understand innovation, it is necessary to accurately measure it. However, innovation, because of its very concept, i.e. the construct of change, is hard to quantify. One could observe innovation through its outcome by measuring the increase in performance after controlling for increases in initial production factors, i.e. by measuring changes in factor productivity (see for example De Loecker, 2007). This is however not always very satisfying, as changes in factor productivity can also occur due to a myriad of other processes, such as changing prices or demands (e.g. see De Loecker, 2011), or advances in management. All of which are not necessarily indicative of technological change.

Innovation thus tends to be measured indirectly through observing the creation and presence of inventions. After all, technological change can only be achieved when new concepts are adopted. Therefore, the definition of innovation can be reformulated to link innovation to invention: innovation is the adoption of invention. Using this observation, the amount of inventions is then a reasonable proxy for the rate of innovation. This is even more so when invention measures can also capture a part of their adoption.

Observing invention is an easier task than directly observing innovation, as inventions lend themselves better for accounting and quantitative measures. Inventions often have technological artefacts, such as products, descriptions and designs. Moreover, inventions can be traded, either directly through intellectual property, or indirectly due to the products which are based on the invention. Using these processes, a value can be attributed to most inventions. Finally, inventions can also be evaluated by their impact on society, amongst others through the value they represent. Therefore, the study of invention allows for a discussion of innovation.

The occurrence of invention can be measured in several ways, for instance by observing the occurrence of new products and designs. This is however cumbersome as new products may not be registered and are often not

categorized in easily accessible data. Therefore, the most popular means of observing inventions is not directly, but rather by the intellectual property, in most cases patents, that protects them (Griliches, 1990). Observing patents, and to a lesser extend trademarks, has the advantage of large, reasonably detailed databases that classify inventions, allowing them to be counted and classified. The main disadvantage is that large numbers of inventions are not patented as they either belong to the category of technology that is not patentable (e.g. services technologies), or they are not patented because there are better alternatives than patenting available (e.g. copy rights, trade secrets, relying on lead time). Therefore, patents statistics should always be viewed as relating to only a subset of the innovative activity in an economy.

Observing innovative efforts through the counting of patents has been practiced since the second part of the 20th century. However, this practice has been criticized for mainly measuring innovation input (e.g. R&D investment), rather than output (Griliches, 1990). The main argument of this critique revolves around the timing of patent applications, which often happen before the final product or process is launched. Therefore, patents may be referring to long abandoned ideas that never reached the stage in which the innovation could be disseminated. Moreover, counting patent applications does not account for the large differences in the value of the different patented inventions, with many patents referring to marginal or unimportant inventions and a few to inventions with an extremely high value. Because of this skewed distribution it is difficult to accurately observe innovation by only counting the number of patents filed. It is for this reason that current scholars use other patent characteristics to qualify the patent counts (for an overview see Squicciarini, 2013). The most popular of these is the number of times a patent has been referenced as relevant prior art by later patents, better known as forward citations or patent citations.

Patents as more than a measure of innovation

Patents represent more than an indirect measure of innovation: patents provide the applicant with a temporary monopoly on producing the described invention. They are therefore important legal instruments in their own right.

The theory behind granting these temporary monopolies rests on the possible market failures in innovation: to create an invention, its prospective owner may need to make considerable expenses. After the invention has been created, it may be easy for competitors to copy the invention.[1] Subsequent competition between the original innovator and its imitators then drives down the price, and consequently the possible profit of the innovator, rendering the original innovator incapable of recouping their expenses that lead to the invention (Arrow, 1962). Rational innovators realize this before they start investing and will refrain from doing so in the first place. Therefore, the failure of the innovator

[1] In practice imitation still requires a substantial investment. Mansfield et al. (1981) estimate that imitation costs in the range of 65% of the original expenses necessary to create the invention in the first place.

to appropriate the benefits of their invention leads to a suboptimal rate of innovation.

Patents counter this process by giving the innovator a temporary monopoly to manufacture and sell their invention. Whenever the innovator detects other actors using the invention protectected by their patent (i.e. infringing)[2], they may seek legal compensation as well as the cessation of the infringing activities. Moreover, through licensing the innovator may also charge others for the privilege of using their invention. Finally, the innovator may also sell the patent to endow others with the full rights of the patent.

Patents have become more than a simple means of protecting an innovative effort: they have now become a part of sophisticated IP strategies that also involve other intellectual property. Patents facilitate markets for technology because they are relatively easy to trade, and they have even become methods of avoiding taxes (Karkinsky and Riedel, 2012).

The uses (and occasional abuses) of patents as intellectual property are still being investigated by many scholars. It is however important to consider that because of this thriving market of Intellectual property (IP), the relation between patent based statistics and information on innovative activities is far from perfect. Therefore, any scholar wishing to use patents needs to be aware of and, if possible, correct for strategic behavior of patent owners.

Patent characteristics as a measure of value

Patents possess a large amount of information due to the need of disclosure: to obtain a patent for an invention, an application needs to be filed with a description detailed enough that a skilled practitioner can recreate it from its description. In practice, patent descriptions are far from clear instructions due to the need for legally precise language, as well as the existence of strategic incentives for the applicant to limit disclosure of their invention.

Nevertheless, patents remain a rich source of information on the inventions they protect. This information can be grouped into several categories. First there is information on the beneficiary of the patent, often referred as the applicant or assignee, as well as information on its inventors. Some offices (e.g. the United States Patent and Trade mark Office (USPTO)) also record the names of the examiners and patent attorneys involved with the patent application. Next there is the description, a text detailing the invention the patent should protect which is accompanied by a set of claims, short texts that outline the legal protection the patent provides. Finally, the patent is summarized in a short abstract.

[2] It is important to note that the infringement doesn't need to be intentional. Firms can be deemed to be infringing without even being aware of the patent they are infringing upon.

During the examination, i.e. the process in which the patent office determines the validity of the patent, more information is added. This involves assigning technological classes that categorize the invention as well as determining relevant prior art, i.e. references to other documents.

Not all information is generally used in innovation research. It is often difficult to create datasets with full descriptions of patents and even then, it is hard to manipulate this information into meaningful economic constructs. Patent abstracts suffer from similar problems but due to their condensed nature and easier accessibility they are becoming more used in research (e.g. Magerman et al. (2010) review textmining methods involving patents). Finally, claims are often counted to identify the value of patents, yet the claim texts themselves are still not used.

The patent characteristics that are often used are the information that ties patents to the owner of the invention, and in fewer instances also the inventors. Other frequently observed characteristics include the technological classes in which the patent is registered which are usefull for industry allocation. They can also be used to determine the patents broadness, i.e. the size of the technological space occupied by the patent. Broader patents are assumed to be more valuable since they likely protect larger inventions. The list of relevant prior art is often used to determine the technological antecedents of the invention, for instance by determining the number of academic sources present as an indication of the scientific contribution of the invention (Callaert et al., 2006; Callaert et al., 2012).

However, most patent qualifiers are determined by not only observing an individual patent but rather by determining the place of the patent in the larger technological space drawn by other patents. For instance, researchers have used combinations of patent classes to find novel patents by identifying patents that combine a set of technological classes for the first time (e.g. Fleming, 2007; Verhoeven et al., 2016).

The most used qualifier stems directly from this idea: forward patent citations. This indicator is observed by counting the number of times a patent is referenced as relevant prior art by other patents. It is an interesting indicator because it represents the relevance attributed to the focal patent by outsiders. Moreover, it presents an interesting analogy to scientific citations, even though this may be deceiving. (Meyer, 2000).

Patent citations as the indicator of patent value

Despite the ever-growing number of patent value indicators, forward patent citations (also known simply as patent citations) remain the most often used indicator to assess the 'value' of a patent. Academic definitions of patent value are not very clear, as they can refer to the private value (Gambardella et al., 2008), the social value (Trajtenberg, 1991; Carpenter et al., 1981), as well as to the knowledge value contained in the innovation (e.g. Hall et al., 2005; Jaffe et al., 1993, 2000; Macgarvie, 2006; Paci and Usai, 2009; Chen, 2017).

Patent citation indicators are widely used due to the availability of the data, and because it is the first quality indicator that has been validated as early as 1981 (Carpenter et al., 1981). Patent citation indicators came into play not only to assess the value of a single patent document, but to estimate the inventive performance of larger entities such as companies (Hall et al., 2005), universities and even countries (Fritch et al., 2014). Patent citation indicators gradually replaced simple patent counts, as the former are deemed to reflect mere inventive input (i.e. R&D expenditure), rather than inventive output (Griliches, 1998).

The use of patent citations in academic studies has increased drastically in recent years (see figure 1). The number of studies that actually uses patent citations to observe patent value has been increasing even more (Jaffe and de Rassenfosse, 2016).

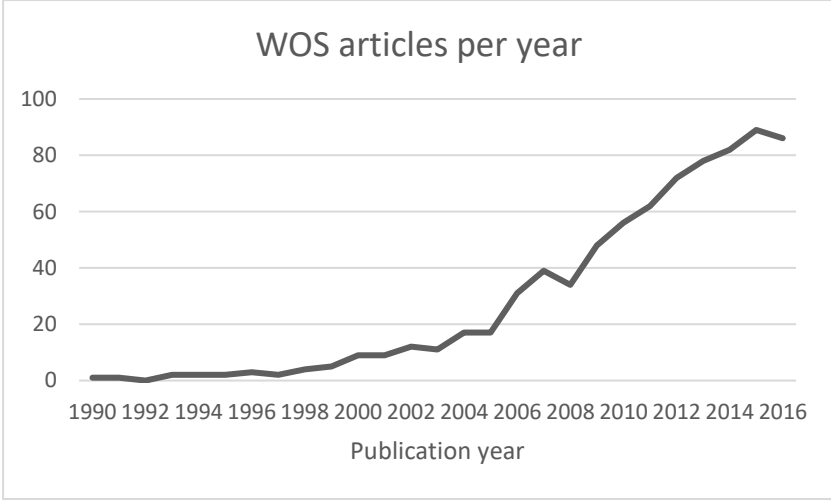

Figure 1: Yearly number of articles that mention "patent citation" in their title or abstract. Data extracted from Web of Science on 17-01-2017.

Currently, patent citations are one of the best validated indicators of patent value, owing to the various validation efforts that have been made. These validation efforts of the patent citation indicator have been done in various ways, employing diverse value constructs as dependent variables and by using different sources of patent citation data. Table 1 provides an overview of important validation efforts, their methods and the results that were found.

Table 1: A selection of scientific works that relate patent citations (in different variations) to (various constructs of) patent value.

Study	Source	Independent Variable	Dependent Variable	Controls	Method	N	Result
Carpenter et al. (1981)	USPTO	Count	R&D award given	Year	2-way ANOVA	202	31.38 F score
Narin and Noma (1987)	USPTO	Count	Financial performance	N/A	Correlation	17	0.628 correlation
Trajtenberg (1990)	USPTO	Aggregated citations	Innovative value on industry level	N/A	Correlation	10	0.685 correlation
Albert et al.(1991)	USPTO	Groups	Relative importance in industry	Year/ company/ technology	ANOVA	77 in 8 groups	2.01 F score (only most cited group significantly different)
Harhoff et al. (1999)	USPTO	Log(count)	Replacement value of patent	none	OLS	192	6.3% of variation explained
Thomas (1999)	USPTO	Count	Renewal decision	ÌPC4, applicant type	Various non parametric analyses	189,359	Significant correlation between patent citations and renewal
Hall et al. (2005)	USPTO	Stock	Tobins Q	R&D, sales, year industry	Non-linear model	12188 (1983 with patents)	3.2% of variation explained
Hall et al. (2007)	EPO	Stock	Tobins Q	R&D, sales, year, industry	Market eq. estimation	1 779	Insignificant
Bessen (2008)	USPTO	Count	Renewal value	Applicant, backward citations. Generality Originality	OLS	48990	4-7% of variation explained
Gambardella et al. (2008)	EPO	Log(count)	Replacement value of patent	Year/country/ tech class[3]	OLS	8 217	1.4% of variation explained
Chen and Chang (2010)	USPTO	Average citations per patent	Market value of pharmaceutical company	Sales, sales growth, other patent portfolio characteristics	Fixed effects OLS	370	Citation indicators are highly significant
Arts et al.(2013)	USPTO	Groups of highly cited	Importance of invention	Year/USPC class	Logistic analyses	74 072	Citation indicators are highly significant

Table 1 shows that patent citations have a significant correlation with patent value: significant results were found in almost all studies reviewed. This is an interesting achievement, considering that so many different dependent variables, as well as many different estimation methods, were used.

The first studies presented in the table, show a large explanatory power, but used smaller samples and considered fewer control variables. Later and larger validation efforts, such as Gambardella et al. (2008) and Hall et al. (2005) provide a more nuanced picture: patent citations are still a significant indicator of economic value but only explain a small part of the variance in the quality indicator, ranging between 2-4% percent. Moreover, most validation studies only pick up significant effects for groups of highly cited patents, as patent citations do not appear to perform well for patents with a lower valuation. Patent

[3] 30 classes of the ISI-INPI-OST classification

citations are therefore an indicator with a consistent but mediocre performance. This was already observed by Gay and Le Bas (2005), and Gittelman (2008).

Recently, there has been growing concern that patent citations are not so consistent as the validation studies in table 1 indicate. For instance, patent citations may be constructed using different sources, and different aggregation techniques. This may substantially alter the performance of patent statistics as they may be biased regarding the country in which the patent originates (Criscuolo, 2006).

Unfortunately, the challenges of the citation indicator are not confined to determining the best method to compute it. It is also important to determine the functional form with which the indicator relates to measures of value. This functional form can deviate from its simplest linear functional form, due to issues such as preferential attachment .in which a patent that is cited more gets an even higher chance of being cited (Hung and Wang, 2010). This poses then two problems: first, the noise in the indicator is related to relatively random occurrences in the time just after the application has been filed; second, this introduces a non-linearity in the relation between citations and quality, which is problematic if we proxy quality using a linear indicator of patent citations. Another reason to deviate from the linear form is the existence of large outliers in the distribution. To prevent these outliers from influencing their analysis, many scholars opt to use a log-linear functional form. There is, however, very little empirical evidence to guide scholars in making the optimal choice.

Furthermore, not all citations are equal. Some scholars choose to incorporate patent citations to family members or equivalents (e.g. Gambardella et al., 2008). The general assumption is that citations made to the equivalents of a patent are of equal stature as those made directly to the patent itself. This assumption may appear reasonable for patents that have the same technical content, i.e. the DOCDB patent family, but becomes more problematic for INPADOC family members that have a looser relation with each other. When multiple data sources are used, scholars can also use patent citations from other offices than the focal office of the application. Since citation practices differ between different patent offices it is not unlikely that a citation from a USPTO patent has a different meaning than one of an EPO patent. Moreover, it is unclear if having a large impact in the European patent system should be equated to having a large impact in the United States patent system.

Finally, as with academic publications, one must also always be vigilant of self-citations. In the case of patents, these are generally defined as citations that come from patents with the same applicant. They are then not an external indication of the impact of the invention but rather an indication that the invention is being used in an ongoing project (Narin et al.,1987), or that subsequent innovation is built on it (Belenzon, 2012). Given the frequent occurrence of measures that correct for the existence of self-citations in scientific publication data, it is remarkable that few studies take patent self-citations into account.

The current literature does not provide enough guidance, by means of large validation exercises, on the concerns of the preceding paragraphs. This lack of validation is detrimental for empirical research, since there may be better alternatives to accepted practices such as the use of different patent citation indicators and the use of the logarithmic transformation. It is also detrimental when it concerns the use of potential newer methods such as using self-citations, or using family based citations.

An overview of this thesis

This doctoral thesis explores the relation between patent citations and patent value. Such an analysis is long overdue due to the large use of patent citations to approximate (several constructs of) patent value on the one hand, and the low correlation between these two concepts on the other. This low correlation indicates that more progress can be made to relate patent citations to patent value. Moreover, the high number of studies that employ patent citations indicates that patent citation research is a general-purpose tool: even small advances here can lead to substantial improvements down the line because many researchers can profit from them.

Three major problems can be identified with the current use of patent citations in innovation research. The first problem concerns the use of patent family based patent indicators. Patent families (i.e. groups of related patents in different offices) could be a better representation of inventions than individual patents. Therefore, patent citations aggregated at this level could be a major improvement in the accuracy of determining patent value. However, there is no research available that evaluated the impact of using these family based indicators. Therefore, in this thesis two chapters will be devoted to this issue. In the first chapter I evaluate the differences that exist between the different patent citation indicators. This chapter shows sufficient grounds for concern, leading to chapter 4, which consists of an evaluation that determines the best patent citation indicator to explain patent value.

The second problem is that of the functional relation between patent citations and patent value. Researchers tend to choose between a linear and a log-linear functional form but neither of these forms have been validated. Therefore, the second chapter presents an evaluation on the benefits of choosing a log-linear over a linear functional form to relate patent citations to patent value.

The third identified problem is that it is still not fully clear why patent citations should correlate with patent value. Most explanations center around the invention that relates to the cited patent, or on the knowledge spillovers that patent citations should represent. However, most of these explanations are based on a lot of assumptions about the relation between patents and innovation. Moreover, they fail to consider the inherent legal character of patents and patent citations. In chapter 3, I therefore advance a new mechanism in which patent citations signal potential litigation, license or acquisition opportunities (i.e. a ‘market value’) for patent holders. This market

value can then be contrasted with the productive value, i.e. the added profits of the patent owner's products due to the monopoly provided by the patent, to arrive at a comprehensive and testable framework that explains decisions by patent owners.

In conclusion, this PHD thesis contributes to our understanding of one of the most used indicators of innovation, and will (hopefully) facilitate better research practices in the future. Below is provided a short overview of the individual chapters of this thesis.

Chapter 1: Patent citations: one size fits all?

Aggregating patent citations in different ways is found to create substantially different indicators, as measured by the correlation structure between them. Moreover, they also identify different 'highly cited patents'. Therefore, favoring one way of calculating a citation indicator over another has non-trivial consequences and, hence, should be given explicit consideration.

This chapter has as co-authors: Dennis Verhoeven, Lin Zhang and Bart Van Looy, and was published in Scientometrics in 2016

Chapter 2: The log-linear relation between patent citations and patent value

In this chapter, I present the results of an analysis concerning patent citation and patent renewal data, thereby advancing a log-linear relation between patent citations and patent value. A complementary analysis related to the patent portfolios of firms, confirms that modelling the relation between patent citations and firm value benefits from the adoption of the log-linear form.

This chapter is single authored and published in Scientometrics (2017)

Chapter 3: Patent citations and a framework of the productive and market value of patents

This chapter introduces a new possible mechanism to relate patent citations to patent value in which patent citations signal potential litigation, license or acquisition opportunities (i.e. a market value) for patent holders. This value is complementary to the productive value of a patent, which is defined as the additional value provided by the patent by protecting the innovative efforts of its owner. This value can be estimated by observing self-citations. After validating that patent citations measure these constructs, a further analysis shows that patent owners tend to value productive value early in the patent life, while market value gains more prominence in later stages.

This chapter has Bart Van Looy as co-author, and parts of this paper were presented at the following conferences: DRUID (2015), DRUID (2016) and Academy of Management (2017).

Chapter 4: Which patent citation indicator performs best at approximating patent value?

This chapter is a follow up to the results found in chapter 1, and it investigates whether using family based patent citation indicators leads to a stronger relation between patent citations and patent value. The results are unfortunately mixed, with patent family indicators only sometimes outperforming the traditional patent citation indicators. Therefore, a deeper investigation is performed, which shows that patent citations to patent family members are mostly valued less than patent citations directly to the patent itself. Moreover, patent citations to certain family members may even indicate competition and crowding out effects. These insights then allow for the creation of a composite indicator, which outperforms any of the more traditionally used indicators.

This chapter has Bart Van Looy as co-author, and parts of this paper were presented at the DRUID (2015) and DRUID (2016) conferences, as part of a composite paper.

References

Albert, M. B., Avery, D., Narin, F., and McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research policy*, *20*(3), 251-259.

Arts, S., Appio, F. P., and Van Looy, B. (2013). Inventions shaping technological trajectories: do existing patent indicators provide a comprehensive picture?. *Scientometrics*, *97*(2), 397-419.

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity: Economic and social factors* (pp. 609-626). Princeton University Press.

Aghion, P., and Howitt, P. (1990). *A model of growth through creative destruction* (No. w3223). National Bureau of Economic Research.

Aghion, P., and Howitt, P. (2007). Capital, innovation, and growth accounting. *Oxford Review of Economic Policy*, *23*(1), 79-93.

Belenzon, S. (2012). Cumulative innovation and market value: evidence from patent citations. *The Economic Journal*, *122*(559), 265-285.

Bessen, J. (2008). The value of US patents by owner and patent characteristics. *Research Policy*, *37*(5), 932-945.

Callaert, J., Van Looy, B., Verbeek, A., Debackere, K., and Thijs, B. (2006). Traces of prior art: An analysis of non-patent references found in patent documents. *Scientometrics*, *69*(1), 3-20.

Callaert, J., Grouwels, J., and Van Looy, B. (2012). Delineating the scientific footprint in technology: Identifying scientific publications within non-patent references. *Scientometrics*, *91*(2), 383-398.

Carpenter, M. P., Narin, F., and Woolf, P. (1981). Citation rates to technologically important patents. *World Patent Information*, *3*(4), 160-163.

Chen, L. (2017). Do patent citations indicate knowledge linkage? The evidence from text similarities between patents and their citations. *Journal of Informetrics*, *11*(1), 63-79.

Chen, Y. S., and Chang, K. C. (2010). The relationship between a firm's patent quality and its market value—the case of US pharmaceutical industry. *Technological Forecasting and Social Change*, *77*(1), 20-33.

Criscuolo, P. (2006). The'home advantage'effect and patent families. A comparison of OECD triadic patents, the USPTO and the EPO. *Scientometrics*, *66*(1), 23-41.

De Loecker, J. (2007). Do exports generate higher productivity? Evidence from Slovenia. Journal of international economics, 73(1), 69-98.

De Loecker, J. (2011). Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. Econometrica, 79(5), 1407-1451.

Fleming, L. (2007). Breakthroughs and the" long tail" of innovation. *MIT Sloan Management Review*, *49*(1), 69.

Frietsch, R., Neuhäusler, P., Jung, T., and Van Looy, B. (2014). Patent indicators for macroeconomic growth—the value of patents estimated by export volume. *Technovation*, *34*(9), 546-558.

Gambardella, A., Harhoff, D., and Verspagen, B. (2008). The value of European patents. *European Management Review*, *5*(2), 69-84.

Gay, C., and Le Bas, C. (2005). Uses without too many abuses of patent citations or the simple economics of patent citations as a measure of value and flows of knowledge. *Economics of Innovation and New Technology*, *14*(5), 333-338.

Gittelman, M. (2008). A note on the value of patents as indicators of innovation: Implications for management research. *The Academy of Management Perspectives*, *22*(3), 21-27.

Griliches, Z. (1990). *Patent statistics as economic indicators: a survey* (No. w3301). National Bureau of Economic Research.

Griliches, Z. (1998). Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence* (pp. 287-343). University of Chicago Press.

Hall, B. H., Jaffe, A., and Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, 16-38.

Hall, B. H., Thoma, G., and Torrisi, S. (2007, August). The market value of patents and R&D: evidence from european firms. In *Academy of Management Proceedings* (Vol. 2007, No. 1, pp. 1-6). Academy of Management.

Harhoff, D., Narin, F., Scherer, F. M., and Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and statistics*, *81*(3), 511-515.

Hung, S. W., and Wang, A. P. (2010). Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network. *Scientometrics*, *82*(1), 121-134.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *the Quarterly journal of Economics*, *108*(3), 577-598.

Jaffe, A. B., Trajtenberg, M., and Fogarty, M. S. (2000). *The meaning of patent citations: Report on the NBER/Case-Western Reserve survey of patentees* (No. w7631). National bureau of economic research.

Jaffe, A. B., and de Rassenfosse, G. (2016). Patent citation data in social science research: overview and best practices.

Karkinsky, T., and Riedel, N. (2012). Corporate taxation and the choice of patent location within multinational firms. *Journal of International Economics*, *88*(1), 176-185.

MacGarvie, M. (2006). Do firms learn from international trade?. *Review of Economics and Statistics*, *88*(1), 46-60.

Magerman, T., Van Looy, B., and Song, X. (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics, 82*(2), 289-306.

Mansfield, E., Schwartz, M., and Wagner, S. (1981). Imitation costs and patents: an empirical study. *The Economic Journal*, *91*(364), 907-918.

Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, *49*(1), 93-123.

Mokyr, J. (1992). *The lever of riches: Technological creativity and economic progress*. Oxford University Press.

Narin, F., Noma, E., and Perry, R. (1987). Patents as indicators of corporate technological strength. *Research policy*, *16*(2-4), 143-155.

Paci, R., and Usai, S. (2009). Knowledge flows across European regions. *The Annals of Regional Science*, *43*(3), 669-690.

Thomas, P. (1999). The effect of technological impact upon patent renewal decisions. *Technology Analysis and Strategic Management*, *11*(2), 181-197.

Schumpeter, J. (1942). Creative destruction. *Capitalism, socialism and democracy*. Harper and Brothers.

Squicciarini, M., Dernis, H., and Criscuolo, C. (2013). Measuring patent quality: Indicators of technological and economic value. *OECD Science, Technology and Industry Working Papers*, *2013*(3)

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand Journal of Economics*, 172-187.

Verhoeven, D., Bakker, J., and Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, *45*(3), 707-723.

Patent citation indicators: One size fits all?

Jurriën Bakker[1], Dennis Verhoeven[1], Lin Zhang[1,2], Bart Van Looy[1]

This chapter has been published in Scientometrics (2016) 106:187–211 DOI 10.1007/s11192-015-1786-0

Abstract

The number of citations that a patent receives is considered an important indicator of the quality and impact of the patent. However, a variety of methods and datasources can be used to calculate this measure. This paper evaluates similarities between citation indicators that differ in terms of (a) the patent office where the focal patent application is filed; (b) whether citations from offices other than that of the application office are considered; and (c) whether the presence of patent families is taken into account. We analyze the correlations between these different indicators and the overlap between patents identified as highly cited by the various measures. Our findings reveal that the citation indicators obtained differ substantially. Favoring one way of calculating a citation indicator over another has non-trivial consequences and, hence, should be given explicit consideration. Correcting for patent families, especially when using a broader definition (INPADOC), provides the most uniform results.

Keywords: Patent citations, EPO, USPTO, PCT, Patent family, Multivariate analysis

JEL Classification O34
Mathematics Subject Classification 62H20 ,62H25, 62H30

Acknowledgments: Lin Zhang acknowledges the support of the National Natural Science Foundation of China, Grant No. 71103064.

[1] Department of Managerial Economics, Strategy and Innovation, KU Leuven, Naamsestraat 69, 3000 Louvain, Belgium

[2] North China University of Water Conservancy and Electric Power, No. 36, Beihuan Road, Zhengzhou 450045, Henan, China

Introduction

The number of times that patents are cited by other patents[1] can be used to complement the mere counting of patented inventions, in order to address the differences in value and impact between inventions. The idea of using patent citations as an indicator is relatively old and appears to have originated from Seidel in 1949 (Karki, 1997). However, the first systematic empirical investigations only emerged in the 1980s, with Carpenter et al. (1981) showing that patents related to industry awards are cited more frequently.

A patent can be cited for various reasons: an inventive step, its industrial relevance, to qualify novelty, or to provide additional relevant information to situate the claims advanced in the patent document. Patents that are cited (more often) are considered more important and valuable than patents that are not used (or used infrequently) to qualify subsequent technological activity. Therefore, one can approximate an individual patent's importance by the number of times it is cited. This argument is empirically supported by the work of Albert et al. (1991), Arts et al. (2012), and Gambardella et al. (2008), who show that patent citations correlate significantly with the value of the individual patent. Likewise, Hall et al. (2005), Narin et al. (1987), Neuhäusler et al. (2011), and Trajtenberg (1990) find a positive correlation between firm performance and the total number of forward citations that their patents receive, even after correcting for firm size. Lanjouw and Schankerman (2004) have determined that patent citations are correlated with other indicators of patent quality, which in turn are correlated with variations in firm value. Additionally, Neuhäusler and Frietsch (2012), and Frietsch et al. (2014) show that forward patent citation counts are strongly correlated with export volume.

While (front page) patent references are ultimately included by examiners, a number of researchers conceive citations as an approximation of knowledge flows: (Hall et al., 2005; Jaffe et al., 1993, 2000; MacGarvie, 2006; Paci and Usai, 2009). When this perspective is adopted, the number of patent citations received indicates the subsequent influence or impact of the knowledge implied in the patented invention.

A major advantage of using patent citations as an indicator of inventive quality, either conceived as value or impact, pertains to the relative simplicity of the measure: it merely requires counting the number of citations a patent receives. Since a large number of patents receive citations[2], this measure allows for the construction of enriched indicators both on the patent level and on more aggregate levels (e.g. firm, industry, country). Currently, patent citations are considered an important indicator of the innovative output of companies (e.g. Hagedoorn and Cloodt, 2003). They also enable statistics and rankings that

[1] Often referred to as patent citations, forward citations or patent citation count. We will use these terms throughout this paper.

[2] Up to 88 % of applications score a non-zero citation count on at least one of the citation indicators we computed.

can be used to determine the innovative performance of countries (e.g. Chakrabarti, 1991; Criscuolo and Verspagen, 2008; Neuhäusler and Frietsch, 2012).

While these, and related studies, point to the relevance of counting the citations received by patent documents, the method of measuring this count is not singularly defined. Despite the simple conceptualization of the measure, calculating citation indicators involves a number of methodological decisions that, in turn, result in a variety of possible citation indicators. The first decision is to choose the data source from which to compile patent citations, given that patent systems are geographically bounded (e.g. US, EU, Japan, China). Since patent citations to one patent system can stem from different geographic areas, the second decision is to choose the source from which citations to the focal set of patents will be included. Finally, given the possible existence of multiple patent documents pertaining to a single invention, a viable option is to treat equivalent patent documents as one patent family, which will also affect citation counts. Currently, there are three different approaches to these decisions in the literature.

The National Bureau of Economic Research (NBER) set up a data platform that contains only patents filed at the United States Patent and Trademark Office (USPTO). This data has been available as early as 2001 (Hall et al., 2001). Additionally, the first analyses on patent citations relied on data from USPTO documents (e.g. Carpenter et al., 1981; Narin et al., 1987). The NBER database is still widely used as the high number of recent citations to the source paper from Hall et al., (2001)[3] attests.

A second set of studies has been conducted using European Patent Office (EPO) patent documents. European patent data is noticeably different from USPTO data: the EPO patents cover a different geographic area; they are heterogeneous in terms of the countries where they are filed; and finally, examiners tend to include fewer citations than their colleagues from the USPTO. Citation data from EPO patents have been compiled since 2003 (Webb et al., 2005) resulting in several EPO-based patent citations studies (e.g. Harhoff and Reitzig, 2004; Neuhäusler et al., 2011; Schoenmakers and Duysters, 2010). Finally, some researchers have opted to go beyond the use of data stemming from a single source (patent office) and take into account the presence of patent families (hence, considering the equivalents of an invention that are present in multiple patent systems when calculating citations). This seems especially appropriate in correcting for 'home biases' (Criscuolo, 2006) and in providing a more encompassing view of the impact of an invention. Examples of this approach can be found in the work of Gambardella et al. (2008), Graham and Harhoff (2006), Magerman et al. (2011), and Neuhäusler and Frietsch (2012).

[3] This paper needs to be cited when the NBER database is used.

When using a patent citation indicator, it is implicitly assumed that different calculation methods of this indicator will, in general, yield similar results. However, this may not necessarily be the case: patent citations from different offices may reflect 'national' impact rather than 'global' impact. Additionally, patent offices focus on their own geographical jurisdiction, which may result in a 'home bias' when looking at patent citations (Criscuolo, 2006). Finally, offices and, hence, examiners' practices vary in terms of the average number of patent citations included: USPTO patent documents display (on average) more citations than EPO patent documents. This, in turn, can lead to a situation whereby citation indicators—derived from different computational choices—do not reflect the same information (Alcácer and Gittelman, 2006). For these reasons, it makes sense to assess the effects of the methodological choices that researchers face when assessing patent quality through forward citations. To the best of our knowledge, no systematic analysis of this kind has been performed. This paper will assess the extent to which different methods yield (dis)similar results. Hence, we pose our research question as follows:

Do citation counts that are computed by different methods reveal similar information?

This question can be further refined by adopting the distinction between technological improvements of an incremental nature vis-à-vis inventions implying a more radical departure from what was previously possible (Baumol, 2004; Dosi, 1982). Accordingly, researchers have operationalized these 'breakthrough' inventions by identifying patents receiving exceptionally high numbers of forward citations (e.g. Ahuja and Lampert, 2001; Chakrabarti, 1991; Schoenmakers and Duysters, 2010). Since citation counts may depend on computational choices, it is of particular interest to compare different methods with respect to identifying highly cited patents. This leads to the following extension of the research question:

To what extent do different calculation methods affect the identification of highly cited patents?

In the remainder of this paper, we answer our first question using correlation and cluster analyses, which compare different methods to calculate citation counts of patent applications. To answer the second research question, we compute the degree of overlap observed between patents that are identified as highly cited by various methods. We start with a systematic discussion of the different computational choices, resulting in a set of indicators that this study then compares. We then present the empirical findings that we obtained and discuss their implications. Overall, our findings signal non-trivial differences among the variety of approaches envisaged.

Overview of the methodological choices when computing patent citation indicators

When counting patent citations, different choices need to be made. These choices pertain to the patent system (or protocol) in which the receiving and citing patent documents reside. The presence of patent families could also be taken into account. In this section, we discuss the general choices that are available when counting forward citations.

The patent office

The patent system in which the patent resides may affect the way in which the patent is cited. This is due to two reasons: the home bias and the inherent difference between the patent systems. A home bias, as discussed in the introduction, implies that patent examiners cite more prior art present in their own jurisdiction[4] (Michel and Bettels, 2001). In addition, while patent systems are largely similar in terms of subject matter and application procedures, they nonetheless differ in several ways. Not only are there observable differences in terms of subject matter - between the USPTO and the EPO concerning the costs incurred (van Pottelsberghe de la Potterie and François, 2009) - but practices such as the 'duty of candor'[5] in the US lead to an increase in references being included in patent documents, which may have an impact on citation-based indicators.

Selection of the citing patents

The second choice a researcher faces relates to selecting the patent documents that cite the focal patent. One can choose to count either the citations that an entity (application or patent family) receives from patents in the same patent office (e.g. EPO, USPTO), or to include citations from patents present in other patent systems. The reason this distinction is worth investigating is twofold.

First, we note that many researchers restrict themselves to a single source, which is often the EPO or the USPTO system, as noted in the introduction. This implies they only count citations that patent applications receive from documents residing in the chosen system. Therefore, it is interesting to examine the effects of this restriction: does restricting citations to the office of the focal application significantly alter the results?

Second, most documents tend, largely, to cite patent documents from within their chosen 'system', due to the examining process (Michel and Bettels, 2001).

[4] We show this later in Table 4.

[5] The 'duty of candor' rule requires that applicant and inventors involved in a patent application must disclose all known information which may adversely affect the probability of obtaining a granted patent.

This is not unexpected since patent examiners should have an overriding concern for the validity of the application within their own jurisdiction. At the same time, when specific procedures are in place, differences can become more pronounced. The case of USPTO is apposite in this respect. When applying to the USPTO, applicants have a so-called duty of candor, requiring them to disclose to the examiner any knowledge of prior art, even if this information could lead to the application being disqualified. Patent examiners then select from these references and/or add other references deemed relevant. However, USPTO examiners are most familiar with USPTO patents. In the case of foreign applicants, references stemming from prior art located outside the American patent system may be advanced relatively more frequently by such applicants. Indeed, Sampat (2004) observed that, in approximately 70 % of patents, references to foreign patents are initially advanced by the applicant (see Azagra-Caro et al., 2011 in this respect).

Correcting for patent families

Patents that represent and/or build on the same invention can also be grouped into so-called 'patent families'. It makes sense to correct citations for the presence of families since other patents can make reference to multiple family members besides the initial, focal application. If the researcher feels that such a citation is just as valuable as a direct citation of the initial patent application, then a correction based on the patent family seems appropriate. In general, this involves adding citations from family members to the citation count of the focal application itself. A case study by Nakamura et al. (2015) shows that accounting for patent families can improve analyses based on patent citations.

There are different definitions of the patent family: in this paper, we consider two. Martínez (2011) defines them as the extended patent family (INPADOC)[6] and the examiner's technology-based family (DOCDB).[7] The DOCDB definition centers on finding the closest equivalents of a patent document in other offices. These documents are usually characterized by having the same priority applications.[8] The INPADOC definition is less strict and is used to find documents protecting the same invention, including documents with a somewhat different priority profile (Albrecht et al., 2010). The members of INPADOC patent families share priority applications with at least one other member of the family. Therefore, patents that are members of the same DOCDB patent family should also be members of the same INPADOC patent family, since all DOCDB patent family members have the same priority

[6] INPADOC is an abbreviation for INternational PAtent DOCumentation, the patent data collected but not generated by the EPO (2014). It is also used to denote the extended patent family in the EPO PATSTAT databases.

[7] DOCDB is the EPO master documentation database (Martínez, 2011). It is also used to denote the examiner's technology-based patent family in the EPO PATSTAT databases.

[8] Albrecht et al. (2010) define the DOCDB patent family as patent applications that have an equal 'priority picture': this can, under certain circumstances, include the priority application itself. Additionally, this family is corrected to include applications that have the same technical content but have been excluded due to a 'discrepancy in the priority picture' Albrecht et al. (2010: 283).

applications.[9] However, it is possible that two members of the same family share no priority applications. This can occur when they both share a priority application with a third member of the family (Lingua, 2005). In this study, both family definitions will be adopted and assessed.

Data and methods

Data used

We used patent data from the October 2011 version of the EPO PATSTAT database. From this data, we extracted indicators for patent applications belonging to the EPO and the USPTO, as well as applications that were filed through the Patent Cooperation Treaty (PCT) route. We chose these applications for two reasons: first, most research that employs patent citation data uses patents from at least one of these three systems (or routes, in the case of PCT applications); second, the data provided by these offices from the USPTO and the EPO is relatively complete in PATSTAT, compared to other offices (also included in PATSTAT). In the remainder of this paper, we shall refer to different origins by designating documents as EPO, USPTO and PCT patent applications.

The focal applications for which the indicators were calculated have been cleaned to remove - amongst others - duplicates caused by untraceable priorities and citations, incorrect conversions of patent numbers, and several issues caused by changes in the USPTO system in 2001.[10] In addition, we only considered USPTO applications that were granted. This is due to the observation that USPTO applications that did not lead to a granted patent are not completely covered by PATSTAT.

After the cleaning exercise, we were left with 8,658,272 focal applications from which 4,397,304 were applications filed at USPTO, 2,343,707 applications filed at EPO and 1,917,261 applications filed via the PCT route. The filing dates range from the 2nd of January, 1970, to the 6th of May 2011. However, it should be noted that the cleaning activity led to the removal of a large number of applications: 3,319,894 applications from the USPTO (mainly because no granted equivalent was yet present); 10,567 applications from the EPO, and 11,335 PCT applications.

[9] This statement holds for the vast majority of patent applications in the EPO PATSTAT database; there is a small minority of patents (0.09 % of DOCDB patent families) that do not fulfill this criterion due to discrepancies in their priority picture. However, these families do not affect the analyses presented later in this paper.

[10] These imply changes in publication types; patent duplicates that occur before and after 2001; and applications that are not available before 2001 but partly available thereafter.

With regard to the citing applications, we used all patent documents available in the 2011 October version of PATSTAT. We excluded only artificial applications.[11] Therefore, the cited applications involved more cleaning than the citing applications. This was carried out because we wanted to keep the citation indicators as close as possible to those obtained when using currently available databases (notably PATSTAT). Consequently, we did not correct all recently known issues that exist in patent citation indicators.[12]

The patent citation indicators and their definitions

We performed four different permutations to calculate our indicators. These are based on patent origin, citation origin and a twofold family correction (see previous section). We have chosen these permutations in the belief that they represent virtually all possible permutations that researchers are likely to consider when working with patent citations. In this section, we explain how these permutations are used.

Starting with patent origin, we compare indicators resulting from three different data sources: EPO, USPTO, and applications filed through the PCT route. We use this data because the vast majority of publications dealing with patent citations use indicators drawn from these sources. Next, we distinguish two groups of indicators based on the source of the citation. This is done by comparing the number of citations received from applications in the office of the focal application, and the number of citations that were received irrespective of the patent office.[13] We will denote those indicators with a restricted source of citations by adding 'within office' to the indicator name.

A third permutation deals with applying a correction for citations received by family members of the focal application. Each family indicator is, therefore, replicated for each patent office. For the patent family definition, we compare both the INPADOC and DOCDB definitions. We denote patent citation indicators that correct for patent family on the cited side (i.e. an indicator that counts all applications that cite the family of the application) by including 'cited family count' in their name. It is possible that a number of citations originate from applications that are part of the same patent family. It can be argued that these citations are mere duplicates since the patent is cited twice by the same invention. This could then create a bias towards citations received from larger patent families, since it is inherent that the size of the family increases the probability of two or more of its members citing the same patent.

[11] These are added to the database to maintain logical links and do not actually represent any patent applications.

[12] An example of this pertains to the well-known issue that EPO references other patents by referring to the references of their PCT equivalents via a non-patent reference in PATSTAT. This has been noted in Harhoff et al. (2003) and Neuhäusler et al. (2011).

[13] In the case of applications filed through the PCT, other applications that followed this route were taken.

Therefore, as a final, fourth permutation, we correct for this bias by counting not the number of patent applications but rather the number of patent families that cite the focal family. We denote patent citation indicators that have this correction by replacing 'cited family count' with 'full family count' in their name. Table 1 provides an overview of the prefixes for the indicators used in this paper.

Table 1: Simplified table of naming indicators. All indicator names consist of a number of prefixes and the word count. This table explains the origins of each prefix. Full definitions for each indicator can be found in Table 2.

Origin of the Prefix	Office of the focal patent	Application or patent family	If patent family correction only applied on the cited side	If patent family correction applied on both sides	If only citations from the office of the focal application are used
Possible prefixes	EPO USPTO PCT	Application DOCDB INPADOC	Cited family	Full family	Within office

This leads to a total of ten different indicators for each office: two indicators based on the application, four indicators based on the DOCDB family, and four indicators based on the INPADOC family. To keep the list of indicators tractable, we provide names and definitions for each indicator in Table 2.

Table 2: Indicators and their definitions. These indicators are calculated for focal applications at the EPO, USPTO and PCT.

Patent Family	Patent citation indicator	Definition
N/A	Simple count	Number of citations a patent application receives from all other patent applications, irrespective of their publication office.
N/A	Simple in count	Number of citations a patent application receives from patent applications which were published in the same office as the focal application.
DOCDB	Family cited	Number of citations the DOCDB patent family of the focal application receives from all other patent applications, irrespective of publication office.
DOCDB	Family in cited	Number of citations the DOCDB patent family of the focal application receives from patent applications which were published in the same office as the focal application.
DOCDB	Full Family count	Number of citations the DOCDB patent family of the focal patent receives from all other DOCDB patent families, irrespective of publication office.
DOCDB	Full Family in count	Number of citations the DOCDB patent family of the focal patent receives from patent applications, which were published in the same office as the focal application. This count is corrected for DOCDB patent family on the citing side.
INPADOC	Family cited	Number of citations the INPADOC patent family of the focal application receives from all other applications, irrespective of publication office.
INPADOC	Family in cited	Number of citations the INPADOC patent family of the focal application receives from other patent applications which were published in the same office as the focal application.
INPADOC	Full Family count	Number of citations the INPADOC patent family of the focal patent receives from all other INPADOC patent families, irrespective of publication office.
INPADOC	Full Family in count	Number of citations the INPADOC patent family of the focal patent receives from other patent applications, which were published in the same office as the focal application. This count is corrected for INPADOC patent family on the citing side.

We computed descriptive statistics for the indicators in Table 2; these are listed in Table 3. From these descriptive statistics, we can derive two main conclusions. The first is that a large number of patents receive at least one citation. However, the rate of patents with a non-zero citation count varies considerably, from 25 % (EPO application count within office) to 88 % (USPTO INPADOC full family count and USPTO INPADOC cited family count). Therefore, the distribution of the citation indicator varies from highly truncated to a more continuous spectrum. Second, we observe that the indicators vary greatly with respect to their averages and standard deviations. The average of the EPO application count within office is about 45 times smaller than the average of the USPTO INPADOC cited family count.

To perform the correlation analysis of the citation indicators, we use only applications that receive at least one citation for any of the indicators considered. In practice, this definition translates into selecting only those applications that receive at least one citation on the DOCDB level or the INPADOC family level. Consequently, other indicators can still have a score of 0. This was done in order to better assess the information contained in the citation counts. Its effects are quite substantial since - depending on the office[14] - a considerable share of patents in our sample have no citations, resulting in identical scores (0) for all indicators. The inclusion of applications that are never cited would have an inflating effect on the correlation and is, therefore, undesirable.

[14] The exact figures are: 21 % for EPO applications, 12 % for USPTO applications, and 37 % for PCT applications.

Table 3: Descriptive statistics for the indicators that were computed for this paper

Focal patent source	Patent Family	Patent citation Indicator	Number of observations	Forward citation statistics			
				Average	Standard deviation	Median	Nonzero
EPO	N/A	Simple count	2,343,707	1.92	5.10	0	38%
EPO	N/A	Simple in count	2,343,707	0.57	1.55	0	25%
EPO	DOCDB	Family cited	2,343,707	9.03	20.88	3	75%
EPO	DOCDB	Family in cited	2,343,707	1.07	2.51	0	41%
EPO	DOCDB	Full Family count	2,343,707	7.28	16.21	3	75%
EPO	DOCDB	Full Family in count	2,343,707	1.03	2.33	0	41%
EPO	INPADOC	Family cited	2,343,707	17.05	84.56	4	79%
EPO	INPADOC	Family in cited	2,343,707	1.76	8.37	0	45%
EPO	INPADOC	Full Family count	2,343,707	11.21	47.77	3	79%
EPO	INPADOC	Full Family in count	2,343,707	1.58	6.43	0	45%
USPTO	N/A	Simple count	4,397,304	9.91	18.22	5	82%
USPTO	N/A	Simple in count	4,397,304	8.46	16.35	4	79%
USPTO	DOCDB	Family cited	4,397,304	13.05	24.66	6	86%
USPTO	DOCDB	Family in cited	4,397,304	10.20	21.08	5	82%
USPTO	DOCDB	Full Family count	4,397,304	10.85	19.48	6	86%
USPTO	DOCDB	Full Family in count	4,397,304	8.97	17.31	4	82%
USPTO	INPADOC	Family cited	4,397,304	25.95	129.50	8	88%
USPTO	INPADOC	Family in cited	4,397,304	19.73	102.23	6	84%
USPTO	INPADOC	Full Family count	4,397,304	16.95	72.90	6	88%
USPTO	INPADOC	Full Family in count	4,397,304	13.54	58.94	5	84%
PCT	N/A	Simple count	1,917,261	1.90	5.63	0	41%
PCT	N/A	Simple in count	1,917,261	0.58	1.55	0	27%
PCT	DOCDB	Family cited	1,917,261	5.73	16.38	1	59%
PCT	DOCDB	Family in cited	1,917,261	1.10	2.49	0	41%
PCT	DOCDB	Full Family count	1,917,261	4.63	12.73	1	59%
PCT	DOCDB	Full Family in count	1,917,261	1.09	2.46	0	41%
PCT	INPADOC	Family cited	1,917,261	13.22	87.36	2	63%
PCT	INPADOC	Family in cited	1,917,261	2.46	15.88	0	46%
PCT	INPADOC	Full Family count	1,917,261	8.63	50.28	1	63%
PCT	INPADOC	Full Family in count	1,917,261	2.31	14.11	0	46%

The distribution of citations

To better understand the behavior of the patent citation indicators, we compiled an overview of the origin and destination of citations, shown in Table 4. This table reveals that the USPTO is the main supplier of citations in the patent system. Not only does the vast majority of citations to USPTO entities come from the USPTO itself, but the USPTO also supplies most citations to other documents. There are more USPTO citations to EPO documents than EPO citations to USPTO documents. A similar pattern emerges for PCT documents.

Correcting for patent family remedies this to some extent at the same time, USPTO documents remain dominant since they account for the most citations overall. In the case of the EPO, INPADOC families with an EPO member receive 6.4 times more citations from USPTO documents than from EPO documents.

Note that the large majority of all citations stem from either USPTO, EPO or PCT documents; very few citations come from other offices such as the Japanese Patent Office (JPO) or the Chinese Patent Office (SIPO). It is interesting to observe that, from the remaining citations, the vast majority are from applications at the national level of the EPO. These citations may indeed represent a duplication of EPO patents, or they may be applications that were filed at only a single national office instead of the EPO, due to the costs of the EPO process - as noted by van Pottelsberghe de la Potterie and François (2009).

Table 4: Origin and destination of citations. Citations are calculated as originating from applications from any office in the PATSTAT database to applications at the EPO, USPTO and PCT. Family correction implies that the citation is made to the patent family of applications at the EPO,USPTO and PCT. The citations are expressed in percentages of all citations to the (patent family of) applications at the focal office.

Family Correction	Focal office	EPO	US PTO	PCT	EPO (National office)[15]	Other	Total	Total citations received
None	EPO	31.35%	36.14%	19.84%	12.31%	0.36%	100%	4,501,136
None	USPTO	4.22%	85.28%	6.62%	3.74%	0.15%	100%	43,566,925
None	PCT	13.30%	31.33%	24.07%	30.72%	0.58%	100%	3,635,340
DOCDB family	EPO	12.03%	64.16%	14.58%	8.85%	0.39%	100%	21,160,972
DOCDB family	USPTO	6.45%	78.18%	8.54%	6.61%	0.22%	100%	57,379,697
DOCDB family	PCT	8.10%	52.14%	16.06%	23.40%	0.30%	100%	10,994,350
INPADOC family	EPO	10.33%	66.25%	15.17%	7.94%	0.31%	100%	39,950,651
INPADOC family	USPTO	6.99%	76.09%	10.69%	6.01%	0.22%	100%	114,120,819
INPADOC family	PCT	7.31%	56.65%	15.93%	19.86%	0.25%	100%	25,338,999

[15] Patent offices that are located in the geographical area that is covered by the EPO.

Patent families

In this paper, we deploy two different family definitions: the DOCDB and the INPADOC definitions. We have compiled some descriptive statistics to understand the effects of correcting for patent family. These statistics are shown in Table 5. Here, we can see that a large number of patent families exist in the database. Note that, even though these families need at least one EPO, USPTO or PCT application, they may also have applications from other offices.

From these patent families, only between 21 and 35 % consist of a single patent application. Most patent families have at least two or more members. Finally, we see that a large number of patent families are equal for either family definition, even after excluding singleton families, which are equal by definition.

Table 5: Statistics of INPADOC and DOCDB families in our applications

Family	Number of families	% Singletons[16]	Average number of members	Overlap between both family definitions	% Overlap[17]	% Overlap[18]
INPADOC	5,309,452	21%	2.64	4,179,052	79%	73%
DOCDB	6,017,825	35%	2.01	4,179,052	69%	63%

[16] Families with only one member
[17] Including singletons
[18] Excluding singletons

Results of the correlation analysis

The effects of expanding the sources of citing patents and correcting for patent family

We first determined the effect of correcting for family and citation origin for each office separately. For this purpose, we compared the 'application count within office' indicator with all other indicators in the office of the focal application. This was done for two reasons: first, the indicator is the most basic (i.e. it is uncorrected for family and only uses citations from its own office); second, it is the indicator that is most widely used - the NBER citation indicator is the USPTO 'application count within office', while the aforementioned scholars who utilize EPO data often use the EPO 'application count within office'. The results of this exercise are presented in Table 6. The full correlation table can be found in ''Appendix 1''.

Table 6: Correlation with the simple in count indicator for each office.
All correlations are significant at the 0.001 level.

Family	Compared indicator	EPO	USPTO	PCT
N/A	Simple count	0.79	0.99	0.77
N/A	Simple in count	1	1	1
DOCDB	Family cited	0.34	0.84	0.35
DOCDB	Family in cited	0.64	0.86	0.72
DOCDB	Full family	0.33	0.84	0.34
DOCDB	Full family in	0.65	0.86	0.72
INPADOC	Family cited	0.09	0.23	0.14
INPADOC	Family in cited	0.20	0.25	0.19
INPADOC	Full family	0.12	0.25	0.16
INPADOC	Full family in	0.26	0.28	0.22

Table 6 shows that there is a substantial effect of citation origin (i.e. all citations vs. only those from within the office) on the patent citation indicators. This can be seen when inspecting the correlation of the 'application count within office' indicator with the 'application count indicator'. This effect is more pronounced for EPO and PCT indicators, with correlations of 0.77–0.79, than for their USPTO equivalent, which is less sensitive in this respect (see the correlation of 0.99). Given the citation information presented in Table 4, this should come as no surprise.

Correcting for patent family introduces considerable differences. The effects of this correction are more outspoken in the EPO and PCT systems than in the USPTO system: where the USPTO 'application count within office' has a correlation of 0.84 with the DOCDB family-corrected indicator, the equivalent correlations for EPO and PCT are situated around 0.33. Correcting for the INPADOC patent family has an even stronger effect than correcting for the DOCDB patent family. Finally, we see that correcting for patent family on the citing side has a relatively small effect. The values in Table 6 are almost equal for the cited family count and the full family count indicators. The tables in ''Appendix 1'' confirm this conclusion: the correlations between cited family count and full family count indicators are very close to 1 for both the DOCDB and the INPADOC family definitions.

The effect of using different sources (for patent documents present in all three systems)

For an inter-office comparison, we calculated the correlation for DOCDB patent families from which applications were filed at the EPO, the USPTO, and through the PCT route. This was done because the DOCDB family is based on the technical equivalence of the documents. Therefore, we can assume that the different elements in the DOCDB family are documents describing the exact same invention in different jurisdictions. Because of this equivalence, a direct comparison focusing on the source document is feasible.

Again, we considered only patents that had at least one citation in their largest (i.e. INPADOC) family. However, we found that all DOCDB patent families with applications in all three offices fulfilled this criterion. Therefore, this restriction did not change the analysis. These considerations led to the comparison of citation indicators for 388,512 DOCDB families. The full correlation matrix is presented in ''Appendix 2''. Here, we extracted the correlations that compare the different sources of patent data. These are listed in Table 7.

Table 7: Correlations between equal indicators derived from different sources. These correlations were calculated on the basis of 388,512 DOCDB families and are significant at the 0.001 level.

	Simple count	Simple in count	DOCDB				INPADOC			
			Family Cited	Family In cited	Full Family	Full Family in	Family Cited	Family In cited	Full Family	Full Family in
EPO-USPTO	0.12	0.09	1	0.71	1	0.75	1.00	0.80	1.00	0.83
EPO-PCT	0.11	0.04	1	0.91	1	0.91	1.00	0.91	1.00	0.91
USPTO-PCT	0.30	0.20	1	0.78	1	0.81	1.00	0.93	1.00	0.95

Table 7 shows that correlations for the basic indicators obtained for the same family but derived from relying on different offices are very low. The correlation between the EPO 'application count within office' and the USPTO 'application count within office' is only 0.09. Using citations from outside the office of the focal application ('application count') remedies this slightly by raising the correlation to levels ranging from 0.11 to 0.30.

Correlations observed when correcting for the DOCDB and INPADOC families are considerably higher. This is naturally the case for the DOCDB cited family count and the DOCDB full family count since the applications are all part of the same family. The INPADOC cited family count and the INPADOC full family count indicators also have coefficients of 1, as shown in Table 7. This is due to the fact that applications that are members of the same DOCDB family are also members of the same INPADOC family. Interestingly, correcting for patent family increases compatibility, even when only citations from the office of the focal application are counted. Therefore, even when there is only application data from one patent office, correcting for the patent family of the focal applications is an interesting method for increasing compatibility with data from other patent offices.

Clustering the patent citation indicators

We performed a cluster analysis on the patent citation indicators by using the correlation table listed in ''Appendix 2'', i.e. pertaining to patent documents that have equivalents in all different systems under study. To define clusters, we performed a divisive cluster analysis, based on factor analysis (see ''Appendix 3'' for a technical description). Since the analysis compares patent applications with the counterparts of their DOCDB patent family, the indicators 'DOCDB cited family count' and the 'DOCDB full family count' give equal values regardless of the office of the focal application. Therefore, they are replaced by the general indicator. This is also carried out for the corresponding INPADOC family indicators since DOCDB family members are also part of the same INPADOC family: the INPADOC family is by definition larger. Including all INPADOC indicators would thus be redundant. The resulting indicators are denoted by the 'ALL' notation. The identified clusters are reported in Table 8.

Table 8: Result of clustering the patent citation indicators.

Source	Family	Indicator	Cluster	R^2 within cluster	R^2 closest Cluster
ALL	INPADOC	CITED	INPADOC	0.9636	0.3586
ALL	INPADOC	FULL	INPADOC	0.9758	0.3918
EPO	INPADOC	Family in cited	INPADOC	0.789	0.7103
EPO	INPADOC	Full Family in count	INPADOC	0.7857	0.7261
PCT	INPADOC	Family in cited	INPADOC	0.9923	0.4306
PCT	INPADOC	Full Family in count	INPADOC	0.9948	0.4448
USPTO	INPADOC	Family in cited	INPADOC	0.9352	0.3164
USPTO	INPADOC	Full Family in count	INPADOC	0.9545	0.3606
EPO	DOCDB	Family in cited	DOCDB B	0.9795	0.4549
EPO	DOCDB	Full Family in count	DOCDB B	0.9816	0.4747
PCT	DOCDB	Family in cited	DOCDB B	0.9808	0.599
PCT	DOCDB	Full Family in count	DOCDB B	0.9805	0.602
PCT	N/A	Simple count	PCT	0.9486	0.203
PCT	N/A	Simple in count	PCT	0.9486	0.2062
USPTO	N/A	Simple count	USPTO	0.9998	0.2108
USPTO	N/A	Simple in count	USPTO	0.9998	0.2041
EPO	N/A	Simple count	EPO	0.9536	0.2817
EPO	N/A	Simple in count	EPO	0.9536	0.2909
ALL	DOCDB	CITED	DOCDB A	0.9891	0.6409
ALL	DOCDB	FULL	DOCDB A	0.9804	0.6734
USPTO	DOCDB	Family in cited	DOCDB A	0.9737	0.5847
USPTO	DOCDB	Full Family in count	DOCDB A	0.993	0.5187

We have created a graphical depiction of the variables and their relation to one another using multidimensional scaling. The result is shown in Fig. 1. The cluster analysis shows that citation indicators that are from different offices (the 'application count' indicators) are significantly different: the corresponding USPTO, EPO and PCT indicators are all grouped into different clusters. This indicates that, when using indicators from USPTO, EPO and PCT sources only, one is relying on different information.

Correcting for patent family substantially increases compatibility. The indicators that are based on the DOCDB family are grouped into only two clusters (clusters DOCDB A and DOCDB B) that appear close to each other (see Fig. 1). It is interesting to note that the USPTO DOCDB family indicators are clustered together with the overall family indicators. This is understandable given the large number of citations that originate from the USPTO system. Finally, we see that the INPADOC indicators are all grouped together in one cluster (cluster INPADOC). Therefore, we conclude that correcting for the INPADOC patent family results in more similar information across patent systems.

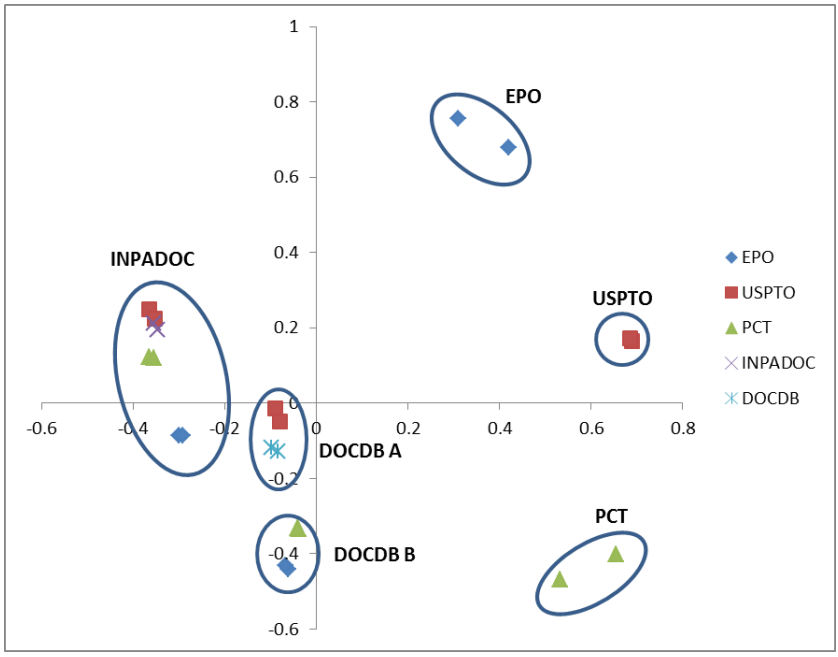

Figure 1: Depiction of the differences between citation indicators on a 2D plane by multidimensional scaling. The dissimilarity between indicators, as defined by $1-R^2$, is represented by the distance between them. Cluster numbers are related to clusters as described in Table 7.

Robustness tests

We performed several robustness tests to verify the results of the correlation analysis under different assumptions and settings. These tests were performed both on the level of the individual sources of the applications (EPO, USPTO and PCT) and the combined set, unless otherwise indicated.

Using a full factor analysis

We performed a full factor analysis on the indicators. We used the principal component method and rotated the solution using the Quartimax algorithm, since this is the most capable method of assigning indicators to different factors. This led to five factors with an eigenvalue larger than 1. We grouped indicators that had loadings higher than 0.5. on the same factor. This analysis resulted in similar conclusions to the cluster analysis: all indicators that relate to patent applications are grouped according to office. However, the family indicators were grouped differently: there was one factor that had all family related indicators, with the exception of the EPO and the PCT DOCDB indicators, which were grouped separately. Thus, a factor analysis groups clusters 1 and 6. We can, therefore, derive the same conclusions as in the cluster analysis: patent citation indicators that relate to equal applications are different from each other, especially when they are related to applications from different patent offices. Family indicators are more similar, but the difference between DOCDB and INPADOC indicators remains present.

Inclusion of uncited applications

In our main analysis, we excluded patent applications that had zero citations on any indicator. This was carried out in order to improve the precision of the analysis. When we included the uncited applications, we found that the correlation of the different indicators increased slightly. However, this increase was small and equally distributed across the different correlation coefficients between the citation indicators. Consequently, we conclude that the inclusion of applications with zero citations does not substantially change the conclusions of the preceding section.

Using only granted applications

The main analysis of the paper pooled different kinds of patent application. It could be that the citation patterns of applications leading to a grant are different from those of other applications. Since granted patent applications are more valuable, researchers could opt to use only those in their analysis. Hence, it is important to determine if our results hold when only considering granted applications.

Patent applications that follow the PCT route cannot be granted (as PCT documents), since the WO is not a patent office with a territory over which it exercises patent grants. Since we only used granted patent applications from the USPTO, the USPTO indicators will not be affected by this step. Therefore, the analysis will only affect the EPO patent applications. For the overall

analysis, we included the PCT and USPTO documents to derive a close comparison with the main analysis.

Using only granted applications from the EPO does not substantially change the correlation between the different indicators. Correlations between indicators on EPO and USPTO documents varied little with the main analysis. This then resulted in the same clusters being returned by the cluster analysis. Nor were the inter-office correlations substantially different. Thus, we conclude that our findings remain similar when including only granted applications.

Using log citations instead of normal citations

Many researchers include not the raw patent citation count but rather the logarithm of the citation count to account for the skewed distribution of patent citations. Therefore, we have also computed the indicators using the following transformation:

$$I^* = \ln(I + 1)$$

whereby I is any of our citation indicators and I^* is its transformed form. We have computed correlations between all transformed indicators. This transformation yields indicators that are more similar to each other. This is because the difference between low and extremely high scores is diminished. Hence, all correlations are substantially improved. This leads the clustering algorithm to select fewer groups. In particular, all DOCDB indicators are now grouped together. All other groups are equal. So, we conclude that, even though the log transformation improves the correlations, this improvement is not sufficient to remove any significant differences that we found in the main analysis.

Using only patent data from before 2000

The main analysis was performed on patent data that cover the time period 1980–2011. Consequently, there are numerous patents that have not yet received (all of their) citations. Since different patent systems may well experience different time lags, this could create a difference in citation data that is due to these time lags, as opposed to an inherent difference in information. In order to control for a potential time lag effect, we repeated the correlation analysis using only patent applications that were filed before 2000.

For our complete analysis, we only compared patent families from which at least one patent in each office had a filing date before 2000. We find that indicators for patents filed before 2000 behave in a similar, albeit not identical, way to the main analysis. The major difference is that the correlations between family-based indicators, most notably those based on INPADOC, increase substantially. This was most pronounced when we computed the full correlation matrix over the three sources of patent data. Because of this, the cluster solution was altered with a reduced number of clusters: one large cluster with all family based indicators, thereby combining clusters INPADOC, DOCDB A and DOCDB B from the main analysis; and three small clusters with application counts from each office, equal to clusters EPO, USPTO and PCT from the main analysis. Consequently, we can conclude that family-based indicators are more similar in this sample, while non-family-based indicators remain very different from each other and from the family-based indicators.

Highly cited patents

Set-up of the analysis

We identified the groups of highly cited patents according to two different criteria: the top 100 patents in terms of citations received, and patents that score more than 5 standard deviations (SD) above the mean number of citations of all patents under study.[19] Highly cited patents were identified, reflecting the unit of analysis of the respective indicators (patent application, DOCDB patent family, INPADOC patent family).

The effects of expanding the sources of citing patents and correcting for patent family

The main observation from the analysis is that commonality between sets of highly cited patents, identified via different indicators, is rather low. This is the case, whether one considers the top 100 cited patents, or whether one considers patents receiving more than 5 standard deviations of citations than average.

Table 9 reports the results obtained in calculating how many identical patent applications are identified when adopting different choices with respect to calculating citations. The reference group consists each time of the patent documents identified by applying the 'application count within office' indicator: citations to the focal document within the patent system of the focal document.

[19] The size of the groups of highly cited patents identified by the 5 SD outlier criterion varies between 765 and 35,145 depending on the source office and indicator specification.

Table 9: Qualified communalities between the simple in count indicator and other indicators from the same office. Fractions are computed as the amount of overlap divided by the maximum amount of possible overlap. Top 100 refers to the 100 most cited patents, and 5 sd. refers to patents present in the 5 standard deviation outlier of the distribution.

Family	Indicator	EPO		USPTO		PCT	
		Top 100	5 sd.	Top 100	5 sd.	Top 100	5 sd.
N/A	Simple count	0.31	0.52	0.89	0.94	0.37	0.52
N/A	Simple in count	1	1	1	1	1	1
DOCDB	Family cited	0.04	0.18	0.76	0.83	0.06	0.16
DOCDB	Family in cited	0.31	0.55	0.81	0.89	0.40	0.54
DOCDB	Full family	0.05	0.18	0.75	0.82	0.06	0.15
DOCDB	Full family in	0.28	0.55	0.80	1.00	0.38	0.54
INPADOC	Family cited	0.04	0.18	0.30	0.72	0.07	0.19
INPADOC	Family in cited	0.20	0.45	0.35	0.78	0.18	0.41
INPADOC	Full family	0.04	0.18	0.28	0.66	0.06	0.17
INPADOC	Full family in	0.20	0.45	0.29	0.71	0.16	0.40

From Table 9, we can derive several conclusions. First, we observe that 5 standard deviation outliers of indicators are in general more similar than the top 100 scores. Second, the table resembles the pattern in Table 8: we observe low levels of overlap for EPO and PCT documents while, for USPTO documents, the overlap is consistently higher. Third, we again observe that both the correction for citation origin and the correction for family have a considerable effect on the indicators. In the case of the EPO and the PCT, we find that the patents identified in the top 100 of the 'application count within office' indicator and those identified by the family corrected indicators hardly overlap.

Even though the commonality improves for the 5 SD outlier and for the USPTO indicators, we conclude that the differences are non-trivial. Differences are larger for INPADOC than for DOCDB indicators.

The effect of using different sources of patent data

In this analysis, we focused on comparing similar indicators from each office with each other. Table 10 presents the result of this analysis. It is important to note that there are two mechanisms by which a highly cited patent does not appear in another patent system. It could be because its family members did not receive a sufficient number of citations, or because it did not have family members present in the other patent system.

In concordance with the results from the previous analysis, we see that using the top 100 rank criterion results in a similar overlap pattern as using the 5 SD outlier criterion. However, the qualified overlap scores are generally lower when using the top 100 rank criterion. Overlaps between indicators that score applications on the citations they receive from within their own offices are very low. This is only slightly improved when citations from other offices are included (moving from 'application count within office' to 'application count' yields, at best, an increase of 3 % for the top 100).

The use of citation indicators that correct for families drastically increases overlap scores between offices. While the use of DOCDB corrected indicators results in qualified overlaps of around 50 %, the highest overlap scores are obtained when INPADOC family corrected citation indicators, which use all citations, are used.

Table 10: Comparison between indicators at different offices. Commonality measures were computed by dividing the number of common members of highly cited groups by the maximum number of common members possible.

Family	Indicator	USPTO – EPO		USPTO – PCT		EPO - PCT	
		Top 100	5 sd. outlier	Top 100	5 sd. outlier	Top 100	5 sd. outlier
N/A	Simple count	0.02	0.09	0.03	0.07	0.00	0.02
N/A	Simple in count	0.02	0.08	0.00	0.03	0.00	0.01
DOCDB	Family cited	0.48	0.72	0.34	0.49	0.57	0.54
DOCDB	Family in cited	0.08	0.21	0.16	0.19	0.14	0.19
DOCDB	Full family	0.45	0.70	0.30	0.46	0.56	0.53
DOCDB	Full family in	0.09	0.24	0.16	0.24	0.14	0.19
INPADOC	Family cited	0.88	0.99	0.78	0.92	0.84	0.74
INPADOC	Family in cited	0.40	0.37	0.43	0.41	0.45	0.33
INPADOC	Full family	0.85	0.99	0.76	0.87	0.85	0.73
INPADOC	Full family in	0.35	0.36	0.43	0.40	0.44	0.34

Conclusion

We set out to determine the (dis)similarity between different citation indicators. We achieved this by computing a set of commonly and less commonly used citation indicators and comparing them with one another. We relied on correlation and cluster analysis to assess (dis)similarities; in addition, we examined which highly cited patents were identified by different indicators. The results showed substantial dissimilarities between the various patent citation indicators.

The correlation and cluster analysis demonstrated that there are large differences in the information revealed by patent citations, depending on which indicator is used. First, a significant effect was present when comparing indicators that use citation information from all offices versus indicators that only use 'within office' citations. Second, indicators computed over different entities (patent application, DOCDB patent family, INPADOC patent family) display only modest levels of commonality. Finally, these effects are most pronounced for EPO and PCT patents. The USPTO indicators tend to be more similar, except when the INPADOC family is corrected for.

Cluster analysis revealed distinctive clusters for each office. Most family corrected indicators, whether they encompass all citations or not, were grouped in clusters reflecting the family definition. Only the indicators based on the DOCDB patent family definition were split into two clusters. Therefore, we conclude that patent citation indicators based on families are more comparable to each other, even when information from only one office is used. This conclusion remains robust under all tests that were performed.

The analysis of highly cited patents provides a similar picture. Correction for the family and the citation origin results in significant effects and leads to larger commonality between different indicators. Commonality is higher when adhering to the indicator reflecting '5 standard deviation' outliers compared to relying on the indicator consisting of the 100 most cited patents. The only indicator resulting in almost complete congruence pertains to the INPADOC corrected indicators.

Since this paper has established clear differences between different citation indicators, it may inspire additional research on the underlying drivers of these differences. Future efforts should be made to examine the origins of these differences. Are they fully explained by different practices in the different offices or do they indicate a separated impact from the regions over which these offices grant patents? A similar effort should be focused on the family indicators. While it appears that they give unbiased information of the global impact of an innovation, this may not be completely true: family indicators

correlate more with USPTO indicators than with their EPO or PCT counterparts. We suggest that this could be due to the higher number of citations that are present in the USPTO system, thus biasing the family indicators towards the greater importance of citation activity in the US. Therefore, efforts could be undertaken to examine the magnitude of this possible bias and, if necessary, derive an unbiased global patent citation indicator. Finally, the INPADOC patent family definition could be further investigated: while the DOCDB definition is clear and often used, this is not the case for the INPADOC patent family definition.

The observation that different indicators display low levels of commonality implies that choices with respect to citation indicators are non-trivial. As a result, we suggest researchers become more aware and explicit in deciding which citation indicator to use. This choice should ultimately be guided by the underlying research question. At the same time, our results may also inspire further research into assessing the consistency of results obtained when deploying different citation indicators. If the intention is to strive for an indicator that is not sensitive to design choices, the INPADOC corrected indicator is clearly the prime candidate since it implies commonality approaching 100 %.

Appendix 1: Correlation between indicators from the same office

Table 11: Correlation of indicators of patents filed at the EPO.

	Family	Indicator	1	2	3	4	5	6	7	8	9	10
1	N/A	Simple count	1									
2	N/A	Simple in count	0.79	1								
3	DOCDB	Family cited	0.40	0.34	1							
4	DOCDB	Family in cited	0.51	0.64	0.66	1						
5	DOCDB	Full family	0.39	0.33	0.99	0.65	1					
6	DOCDB	Full family in	0.52	0.65	0.66	0.99	0.65	1				
7	INPADOC	Family cited	0.12	0.09	0.36	0.26	0.35	0.25	1			
8	INPADOC	Family in cited	0.17	0.20	0.28	0.39	0.28	0.38	0.88	1		
9	INPADOC	Full family	0.14	0.12	0.40	0.30	0.40	0.30	0.91	0.77	1	
10	INPADOC	Full family in	0.23	0.26	0.35	0.47	0.35	0.48	0.81	0.89	0.87	1

Table 12: Correlation of indicators of patents filed at the USPTO

	Family	Indicator	1	2	3	4	5	6	7	8	9	10
1	N/A	Simple count	1									
2	N/A	Simple in count	0.99	1								
3	DOCDB	Family cited	0.85	0.84	1							
4	DOCDB	Family in cited	0.85	0.86	0.99	1						
5	DOCDB	Full family	0.84	0.84	0.99	0.98	1					
6	DOCDB	Full family in	0.85	0.86	0.98	0.99	0.99	1				
7	INPADOC	Family cited	0.23	0.23	0.30	0.31	0.30	0.30	1			
8	INPADOC	Family in cited	0.25	0.25	0.31	0.32	0.31	0.32	0.99	1		
9	INPADOC	Full family	0.25	0.25	0.33	0.33	0.33	0.33	0.95	0.94	1	
10	INPADOC	Full family in	0.27	0.28	0.34	0.35	0.35	0.35	0.94	0.95	0.99	1

Table 13: Correlation of indicators of patents filed at the PCT

	Family	Indicator	1	2	3	4	5	6	7	8	9	10
1	N/A	Simple count	1									
2	N/A	Simple in count	0.77	1								
3	DOCDB	Family cited	0.52	0.35	1							
4	DOCDB	Family in cited	0.61	0.72	0.69	1						
5	DOCDB	Full family	0.49	0.34	0.99	0.68	1					
6	DOCDB	Full family in	0.61	0.72	0.69	1.00	0.68	1				
7	INPADOC	Family cited	0.21	0.14	0.29	0.23	0.29	0.23	1			
8	INPADOC	Family in cited	0.20	0.19	0.20	0.28	0.20	0.28	0.88	1		
9	INPADOC	Full family	0.22	0.16	0.32	0.26	0.32	0.26	0.93	0.82	1	
10	INPADOC	Full family in	0.22	0.22	0.23	0.31	0.23	0.31	0.84	0.94	0.88	1

Appendix 2: Correlation between indicators from different offices

Table 14: Correlation coefficients of indicators pertaining to patents filed both at the EPO (columns) and the USPTO (rows)

	Family	Indicator	1	2	3	4	5	6
1	N/A	Simple count	0.12	0.09	0.27	0.28	0.17	0.18
2	N/A	Simple in count	0.12	0.09	0.27	0.27	0.17	0.18
3	DOCDB	Family in cited	0.11	0.06	0.71	0.72	0.81	0.82
4	DOCDB	Full Family in count	0.12	0.07	0.74	0.75	0.82	0.83
5	INPADOC	Family in cited	0.01	0.00	0.39	0.40	0.80	0.80
6	INPADOC	Full Family in count	0.02	0.00	0.43	0.45	0.82	0.83

Table 15: Correlation coefficients of indicators pertaining to patents filed both at the EPO (columns) and the PCT (rows)

	Family	Indicator	1	2	3	4	5	6
1	N/A	Simple count	0.11	0.07	0.47	0.47	0.33	0.33
2	N/A	Simple in count	0.07	0.04	0.34	0.33	0.22	0.22
3	DOCDB	Family in cited	0.16	0.10	0.91	0.91	0.82	0.82
4	DOCDB	Full Family in count	0.16	0.10	0.91	0.91	0.82	0.82
5	INPADOC	Family in cited	0.03	0.01	0.54	0.55	0.91	0.91
6	INPADOC	Full Family in count	0.04	0.01	0.55	0.56	0.92	0.91

Table 16: Correlation coefficients of indicators pertaining to patents filed both at the USPTO (columns) and the PCT (rows)

	Family	Indicator	1	2	3	4	5	6
1	N/A	Simple count	0.30	0.29	0.37	0.38	0.11	0.13
2	N/A	Simple in count	0.20	0.19	0.22	0.23	0.04	0.06
3	DOCDB	Family in cited	0.31	0.30	0.78	0.81	0.48	0.52
4	DOCDB	Full Family in count	0.31	0.30	0.78	0.81	0.48	0.52
5	INPADOC	Family in cited	0.10	0.10	0.78	0.76	0.93	0.94
6	INPADOC	Full Family in count	0.10	0.11	0.79	0.78	0.94	0.95

Appendix 3: Variable cluster method

This appendix explains the cluster algorithm that was used to cluster indicators. This method is an implementation of the VARCLUS procedure in the SAS software package (SAS Institute, 2008). What follows are excerpts from the SAS manual (SAS Institute, 2008: 7461–7463) explaining the logic of the underlying procedure. Our specific settings are detailed in italics. Options not related to our analysis have been omitted.

'The VARCLUS procedure divides a set of numeric variables into disjoint or hierarchical clusters. Associated with each cluster is a linear combination of the variables in the cluster. The linear combination used here consists of the first principal component. (…) The first principal component is a weighted average of the variables that explains as much variance as possible.

(…) The VARCLUS procedure tries to maximize the variance that is explained by the cluster components, summed over all the clusters. The cluster components are oblique, not orthogonal, even when the cluster components are first principal components. In an ordinary principal component analysis, all components are computed from the same variables, and the first principal component is orthogonal to the second principal component and to every other principal component. In the VARCLUS procedure, each cluster component is computed from a different set of variables than all the other cluster components. The first principal component of one cluster might be correlated with the first principal component of another cluster. Hence, the VARCLUS algorithm is a type of oblique component analysis.

We use the correlation matrices as input for the principal component analysis used in the VARCLUS procedure (…) The VARCLUS algorithm is both divisive and iterative. By default, the VARCLUS procedure begins with all variables in a single cluster. It then repeats the following steps:

1. A cluster is chosen for splitting. Depending on (…) the largest eigenvalue associated with the second principal component (…)

2. The chosen cluster is split into two clusters by finding the first two principal components, performing an orthoblique rotation (raw quartimax rotation on the eigenvectors; Harris and Kaiser 1964), and assigning each variable to the rotated component with which it has the higher squared correlation.

3. Variables are iteratively reassigned to clusters to try to maximize the variance accounted for by the cluster components.

(…) VARCLUS stops splitting when every cluster has only one eigenvalue greater than one, thus satisfying the most popular criterion for determining the sufficiency of a single underlying dimension.'

References

Ahuja, G., and Lampert, C. Morris. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, 22(6–7),521–543.

Albert, M. B., Avery, D., Narin, F., and McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20(3), 251–259.

Albrecht, M. A., Bosma, R., van Dinter, T., Ernst, J. L., van Ginkel, K., and Versloot-Spoelstra, F. (2010). Quality assurance in the EPO patent information resource. *World Patent Information*, 32(4), 279–286.

Alcácer, J., and Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, 88(4), 774–779.

Arts, S., Appio, F., and Van Looy, B. (2012). Inventions shaping technological trajectories: Do existing patent indicators provide a comprehensive picture? *Scientometrics*, 97(2), 397–419.

Azagra-Caro, J. M., Mattsson, P., and Perruchas, F. (2011). Smoothing the lies: The distinctive effects of patent characteristics on examiner and applicant citations. *Journal of the American Society forInformation Science and Technology*, 62(9), 1727–1740.

Baumol, W. J. (2004). Education for innovation: entrepreneurial breakthroughs vs. corporate incremental improvements (No. w10578). *National Bureau of Economic Research*.

Carpenter, M. P., Narin, F., and Woolf, P. (1981). Citation rates to technologically important patents. *World Patent Information*, 3(4), 160–163.

Chakrabarti, A. K. (1991). Competition in high technology: Analysis of patents of US, Japan, UK, France, West Germany, and Canada. *Engineering Management, IEEE Transactions on*, 38(1), 78–84.

Criscuolo, P. (2006). The 'home advantage' effect and patent families. A comparison of OECD triadic patents, the USPTO and the EPO. *Scientometrics*, 66(1), 23–41.

Criscuolo, P., and Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs.examiner citations in European patents. *Research Policy*, 37(10), 1892–1908.

Dosi, G. (1982). Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change. *Research policy*, 11(3), 147–162.

EPO. (2014). What's in a name? *Patent Information News*, 4, 4.

Frietsch, R., Neuhäusler, P., Jung, T., and Van Looy, B. (2014). Patent indicators for macroeconomic growth—The value of patents estimated by export volume. *Technovation*, 34(9), 546–558.

Gambardella, A., Harhoff, D., and Verspagen, B. (2008). The value of European patents. *European Management Review*, 5(2), 69–84.

Graham, S., and Harhoff, D. (2006). Can post-grant reviews improve patent system design? A twin study of European and US patents. *CEPR Discussion Paper No. 5680*, CEPR London.

Hagedoorn, J., and Cloodt, M. (2003). Measuring innovative performance: Is there an advantage in using multiple indicators? *Research Policy*, 32(8), 1365–1379.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools (No. w8498). *National Bureau of Economic Research*.

Hall, B. H., Jaffe, A. B, and Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics*,36(1), 16–38.

Harhoff, D., and Reitzig, M. (2004). Determinants of opposition against EPO patent grants—The case of biotechnology and pharmaceuticals. *International Journal of Industrial Organization*, 22(4), 443–480.

Harhoff, D., Scherer, F. M., and Vopel, K. (2003). Citations, family size, opposition and the value of patente rights. *Research Policy*, 32(8), 1343-1363.

Harris, C. W., and Kaiser, H. F. (1964). Oblique factor analytic solutions by orthogonal transformations.*Psychometrika*, 29(4), 347–362.

Jaffe, A. B., Trajtenberg, M., and Fogarty, M. S. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *The American Economic Review*, 90(2), 215–218.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics*, 108(3), 577–598.

Karki, M. M. S. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information*, 19(4), 269–272.

Lanjouw, J. O., and Schankerman, M. (2004). Patent quality and research productivity: Measuring innovation with multiple indicators. *The Economic Journal*, 114(495), 441–465.

Lingua, D. G. (2005). INPADOC: 30 years of endeavours yet unmapped territories remain! *World Patent Information*, 27(2), 105–111.

MacGarvie, M. (2006). Do firms learn from international trade? *Review of Economics and Statistics*, 88(1), 46–60.

Magerman, T., Van Looy, B., and Debackere, K. (2011). In search of anticommons: Patent-paper pairs in biotechnology. An analysis of citation flows. *MSI FEB Working paper*, KU Leuven.

Martínez, C. (2011). Patent families: When do different definitions really matter? *Scientometrics*, 86(1),39–63.

Michel, J., and Bettels, B. (2001). Patent citation analysis. A closer look at the basic input data from patente search reports. *Scientometrics*, 51(1), 185–201.

Nakamura, H., Suzuki, S., Kajikawa, Y., and Osawa, M. (2015). The effect of patent family information in patent citation network analysis: A comparative case study in the drivetrain domain. *Scientometrics*, 104(2), 437–452.

Narin, F., Noma, E., and Perry, R. (1987). Patents as indicators of corporate technological strength. *Research Policy*, 16(2), 143–155.

Neuhäusler, P., and Frietsch, R. (2012). Patent families as macro level patent value indicators: Applying weights to account for market differences. *Scientometrics*, 96(1), 1–23.

Neuhäusler, P., Frietsch, R., Schubert, T., and Blind, K. (2011). Patents and the financial performance of firms—An analysis based on stock market data (No. 28). *Fraunhofer ISI discussion papers innovation systems and policy analysis*.

Paci, R., and Usai, S. (2009). Knowledge flows across European regions. *The Annals of Regional Science*, 43(3), 669–690.

Sampat, B. N. (2004). Examining patent examination: an analysis of examiner and applicant generated prior art (Doctoral dissertation, University of Michigan).

SAS Institute Inc. (2008). SAS/STAT_ 9.2 User's Guide. Cary, NC: SAS Institute Inc.

Schoenmakers, W., and Duysters, G. (2010). The technological origins of radical inventions. *Research Policy*, 39(8), 1051–1059.

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The RAND Journal of Economics*,12(1), 172–187.

van Pottelsberghe de la Potterie, B., and François, D. (2009). The cost factor in patent systems. *Journal of Industry, Competition and Trade*, 9(4), 329–355.

Webb, C., Dernis, H., Harhoff, D., and Hoisl, K. (2005). Analysing European and international patent citations: A set of EPO patent database building blocks, *OCDE Science*. Technology and Industry Working Paper, 9.

The log-linear relation between patent citations and patent value

Jurriën Bakker[1]

This chapter has been published in Scientometrics (2017) 110:879–892 DOI 10.1007/s11192-016-2208-7

Abstract

This paper reports the results of an analysis of patent citation and patent renewal data, advancing a log-linear relation between patent citations and patent value. A complementary analysis of firms' patent portfolios confirms that modelling the relation between citations and firm value benefits from the adoption of the log-linear form.

Keywords: Patent citations, Patent value, Patent renewal, Tobin's Q

JEL Classification O34
Mathematics Subject Classification 62N68, 91B84

Acknowledgements: The author wishes to thank Bart Van Looy, Otto Toivanen, Dirk Czarnitzki, and two anonymous reviewers for their insightful comments and invaluable support in completing this paper.

1 Department of Managerial Economics, Strategy and Innovation, KU Leuven, Naamsestraat 69, 3000 Louvain, Belgium.

Introduction

Forward patent citations are a ubiquitous indicator of the impact and value of patents and, by extension, patent portfolios. Forward citations - i.e. the number of times a patent is deemed relevant prior art by the examiners and/or applicants of later patents - have obtained this position in part due to the extensive validation of this indicator. Validation efforts progressed from early small-scale studies (Carpenter et al., 1981; Trajtenberg, 1990) to larger studies involving large patent sets (e.g. Bessen, 2008; Hall et al., 2005; Gambardella et al., 2008). Moreover, patent citations have been found to correspond to several constructs of value: innovative value (Albert et al., 1991; Arts et al., 2013; Carpenter et al., 1981; Trajtenberg, 1990), private value (Harhoff et al., 1999, 2003; Gambardella et al., 2008) and market value (Belenzon, 2012; Hall et al., 2005).

However, several authors have signaled a disturbing lack of explanatory power when using patent citations to explain patent value (e.g. Gay and Le Bas, 2005; Gittelman, 2008). Furthermore, the distribution of patent citations is skewed and often involves outliers. Consequently, log-linear transformations have been advanced as a solution (e.g. Harhoff et al., 1999; Gambardella et al., 2008). Moreover, a non-linear approach may be required because of the specificities of the patent citation network. For instance, Hung and Wang (2010) found that patent citations follow a rule of preferential attachment. This phenomenon, as first outlined by Barabási and Albert (1999), entails in this case that patents more frequently cite patents that have already been cited, regardless of quality considerations. Therefore, later citations could have a lesser value, thus indicating that patent citations do not scale linearly with patent value.

In most studies that involve patent data, patents statistics are grouped. This can be undertaken on the level of a firm's patent portfolio (e.g. Hall et al., 2005), or on national levels (e.g. Neuhäusler and Frietsch, 2012). If patent value does not scale linearly with the sum of patent citations on the individual level, this would also have implications for the group level. Consequently, patent portfolios would have to be calculated differently; simply taking the sum of patent citations to patents in the portfolio would be inadequate. The continued relevance of patent citations as an important measure of value, as noted in Jaffe and de Rassenfosse (2016), increases the importance of better understanding the relation between patent citations and patent value. This will help in analyses where patent value is modeled as a dependent variable, used as an independent variable, or used as a control. Improvements in using patent citations as a proxy for patent value may also benefit research on patent portfolios and, by extension, modeling the innovative performance of large actors such as firms and countries. Moreover, it secures a better insight into the processes by which patent citations have come to be correlated with patent value.

In this paper, the relevance of a log-linear relationship is demonstrated by relating patent citations to patent renewal data. This is undertaken for patents from both the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO), because they have different renewal characteristics. Consequently, comparing the analyses of patents from these offices will provide robustness to the results. In a complementary analysis, the derived functional form is assessed by analyzing the value of firm portfolios in an adapted replication of Hall et al. (2005). The results reveal a substantial increase in explanatory power that may occur when adopting a model specification that reflects the log-linear relationship.

Measuring the relation between citations and renewal

In this analysis, the relationship between patents citations and patent value - as indicated by the decision of patent owners to pay maintenance fees (i.e. patent renewal) - is assessed. Patent renewal can be considered an indicator of private patent value (Pakes and Schankerman, 1984; Pakes, 1986; Lanjouw et al., 1998; Harhoff et al., 1999; Thomas, 1999; Hegde and Sampat, 2009; Van Pottelsberghe de la Potterie and Van Zeebroeck, 2008; de Rassenfosse and Van Pottelsberghe de la Potterie, 2013; de Rassenfosse and Jaffe, 2014), since renewal reflects an economic decision on the part of the patent owner. In other words, it registers a minimal private value that the owner assigns to the patent.

Patent data is obtained from the European Patent Office (EPO) PATSTAT fall 2013 database, and complemented with renewal data as observed in fee payments to the relevant patent office, from its spring 2014 counterpart. The sample was constructed to allow a comparison between EPO and United States Patent and Trademark Office (USPTO) patent applications. Therefore, only DOCDB[20] patent families with granted applications at both the EPO and the USPTO have been included. Using patent applications from 1981 to 2000, the sample includes 547365 granted EPO applications and 571816 granted USPTO applications. To better allow a comparison between results obtained for the EPO patents and the USPTO patents, patent citations are observed as citations made to the DOCDB family of the patent by other DOCDB patent families. This measure is comparable across patent offices, unlike the counts of citations to individual patents of different offices, which are affected by different citation practices practiced in different offices and differ considerably (Bakker et al., 2016).

[20] This family groups patents from different offices that have an identical technical content (Albrecht et al., 2010).

Observing patent renewal for USPTO patents is relatively easy, as one simply has to observe whether maintenance fees have been paid to the USPTO. The USPTO system anticipates three periods of renewal with decisions possible at 4, 8 and 12 years of the patent life. The USPTO renewal time is calculated as the period for which fees have been paid.

Observing patent renewal for EPO patents is more complicated because the EPO has no unitary structure but acts as an intergovernmental organization operating through member state offices. Granted patents are hosted at national offices that subscribe to the EPO (e.g. the Portuguese or the Netherlands patent office). Maintenance payments and renewal decisions are also made at these offices. Thus, an EPO patent may be renewed at one office but abandoned at another. In order to achieve a single EPO renewal indicator, the Single Renewal Approach (SRA) of Van Zeebroeck (2011) is used: the EPO renewal indicator is determined by the longest time a patent has been renewed at any of the national offices subscribing to the EPO convention. Renewal payments at the national offices that subscribe to the EPO are made yearly, and EPO renewal time is therefore calculated as the longest period for which fees have been paid at any of these national offices.

An analysis is performed where a dummy is created for each score level of the DOCDB citation indicator. Here, the dummies are denoted as $DOCDB_i$, where i denotes the citation score. Levels range from 0, 1, 2, 3…99, 100…368+. Patents with a score larger than 100 are grouped in progressively larger clusters consisting of not one, but several, levels of the citation score. For example, patents with a DOCDB citation score of 101, 102 or 103 are grouped in the same cluster. These clusters are treated in the same way as individual citation levels, where their citation score is determined by the central value of the patent citation score of the levels grouped in each cluster. This procedure is undertaken because the density of patents per citation level is otherwise too low for meaningful estimation of the coefficients associated with each level. The number of citation levels per cluster is denoted in Table 1. Finally, all patents with a score of 368 and above (i.e. 9 standard deviation outliers) are grouped together in one category. Because there are few uncited patents, the reference category combines the set of uncited patents with the set where patents are cited only once.

Table 2: Number of citation levels grouped together in each cluster as a function of the DOCDB citation score.

DOCDB citation score	Citation levels per cluster
2-100	1
101-142	3
143-178	5
179-227	7
228-290	9
291-367	11
368+	N/A

Including a set of appropriate control variables ($x_{controls}$) and assuming an independent error term ε, the number of years the patent was maintained ($t_{renewal}$) can be expressed as a function $f()$ of the citation levels $DOCDB_i$ and a constant C:

$$t_{renewal} = f(C + \sum_i \beta_i \cdot DOCDB_i + \sum_{controls} \beta_{controls} \cdot x_{controls}) + \varepsilon$$

In this model, assuming the function $f()$ is correctly chosen to represent the relation between patent value and renewal time, the size of the coefficients β_i should relate to i following the functional form with which patent citations relate to patent value. Pakes (1986) highlighted a real option approach to the estimation of renewal time by considering that patent renewal not only extends patent protection for a limited time but also provides the option of future extensions. Maurseth (2005) modeled this as a survival problem using a Cox model. This approach rests on the idea that an expected revenue stream can be attributed to a patent in each given year. Whenever the costs of maintaining the patent (are expected to) exceed the revenue stream, the owner of the patent will decide not to continue paying maintenance fees. Because the (modeled) revenue and the costs of maintaining a patent are not constant, the relation between patent value and observed patent life is not linear. Therefore, a Cox survival analysis should better model patent value through patent renewal than a linear regression model such as OLS. The survival model can also take into account the censoring that stems from either the data that is absent due to missing renewal information at the end of the dataset or from the maximum patent lifetime of 20 years.

Cox survival regressions are dependent on the number of distinct possible survival times that can be observed. Multiple objects with the same survival time need to be taken into consideration and a method needs to be employed to resolve these ties. This is especially important for USPTO cases where only three renewal decisions are taken for each patent, resulting in many patents with the exact same survival time. In analyses with many ties, the standard method of resolving them (Breslow, 1974) could yield biased coefficients, while the Efron (1977) method has been advocated as an unbiased method (Hertz-Picciotto and Rockhill, 1997; Hsieh, 1995). Therefore, the analyses have been computed using both methods. Small differences were found, but these were not significant in estimating the log-linear fits presented in the results section. In the main analyses, the results of Efron's method are reported, since they appear to be less biased than those of Breslow's method (Hertz-Picciotto and Rockhill, 1997).

This paper attempts to construct a framework that applies to all patents. The analysis, therefore, includes control variables concerning the year and the technological class (IPC3 level) of the application, because of the likelihood that these variables affect patent citations as well as renewal probability. Furthermore, it is likely that different applicants have different renewal considerations and write different patents, resulting in different citation characteristics. Thus, the analysis also includes controls that reflect basic

characteristics of the applicant (i.e. type, experience, size and country[21]). This information has been obtained from the harmonized table provided for the EPO PATSTAT database (Magerman et al., 2006; Peeters et al., 2010). Table 2 provides an overview of the definitions and descriptive statistics of the variables used in the analysis.

Table 3: Descriptions and descriptive statistics of USPTO patents in the Cox survival analyses. Statistics for EPO patents deviate slightly and are given for EPO renewal. [a] In the case of dummy variables relating to levels of a discrete variable, statistics are given for this variable. [b] indicates partial counts when applicable. Finally, when an application is co-patented, variables with [c] default to the largest and oldest applicants.

Name	Description	Mean	Standard deviation	Min	Max
USPTO renewal time	Maximum year for which maintenance fees are paid at the USPTO.	14.59	5.76	4	20
EPO renewal time	Maximum year for which maintenance fees are paid at any national office subscribing to the EPO.	13.12	4.81	2	20
$DOCDB_i$[a]	Dummy indicating whether the DOCDB patent family of the patent is cited *i* times by other DOCDB families.	22.16	39.02	0.00	3146
Nr. Offices	Number of distinct patent offices in which the DOCDB family of the patent has at least 1 application present.	7.08	4.12	2	51
Application year[a]	Dummy for the application year of the patent.	1992.59	5.38	1981	2000
IPC3[b]	Dummy variable to indicate if the IPC3 class (e.g. A01) is present in the patent application.	N/A	N/A	N/A	N/A
Applicant experience[c]	Years between filing of current patent and that of the first application filed by the applicant.	36.23	31.12	0	146
Ln(Applt. size)[c]	Logarithm of the total number of patents filed by the applicant.	7.40	3.32	0	12.99
Co-patented	Dummy indicating if the patent has more than 1 applicant.	0.06	0.24	0	1
Applicant type[b]	Type of applicant: company, government, hospital, individual, university or unknown.	N/A	N/A	N/A	N/A
Applicant country[b]	Dummy for the country in which the applicant resided at time of filing the patent.	N/A	N/A	N/A	N/A

[21] Applicants may come from countries with few patents, which would disrupt the analysis because the dummy variable relating to that country cannot be estimated (well). Therefore, applicant countries with less than 50 patents in the analyses have been grouped together in a separate category. This has affected 819 patents in total.

Results

The coefficient estimates (β_i) for the $DOCDB_i$ dummies, from the Cox survival regression for the USPTO renewal data, are shown in Fig. 1. Note that Cox survival regressions estimate hazard (i.e. abandoning patents); hence, negative coefficients indicate a higher chance of renewal. The coefficients of the analysis are monotonically decreasing, even though, at higher citation levels, there is greater variance in this relation.

A log-linear function of the form $\beta_i = a + b\ln(c + i)$ is estimated with β_i the size of the dummy coefficient, and i the citation score. The log-linear relation depicted fits the relation between the citation scores and the coefficients well (R^2 = 0.86). As a comparison, a linear curve of the form $\beta_i = a + bi$ has also been estimated, which produces a worse fit (R^2 = 0.54). Finally, it can be argued that the log-linear fit has one more parameter and would, therefore, have an advantage over the linear estimation. Consequently, a quadratic curve of the form $\beta_i = a + bi + ci^2$ has been estimated as well, and it is a better fit than the linear specification (R^2 = 0.76). Nevertheless, the log-linear specification is a superior fit to this specification, indicating that the data fit better with a log-linear form than a polynomial with the same number of parameters.

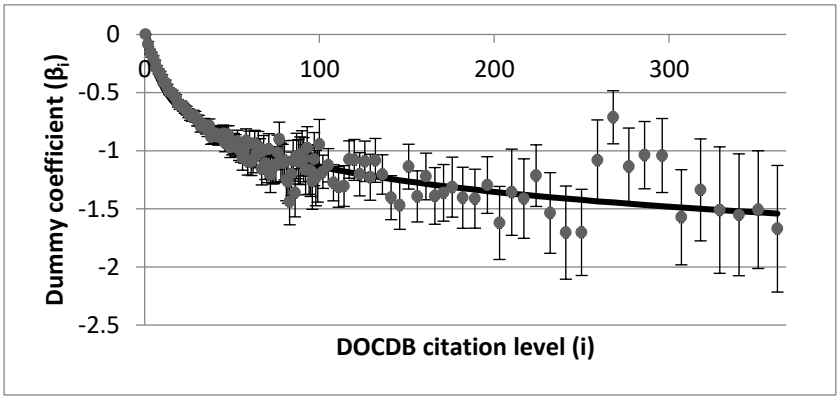

Figure 2: Estimates of the dummy coefficients related to different DOCDB citation scores, that were obtained from a Cox survival analysis relating different scores on the DOCDB citation indicator, to the maintenance time of a patent at the USPTO. A 95% confidence interval is shown as well as a log-linear fit of $\beta_i = 0.33 - 0.32\ln(1.43 + i)$, which has an R^2 of 0.86.

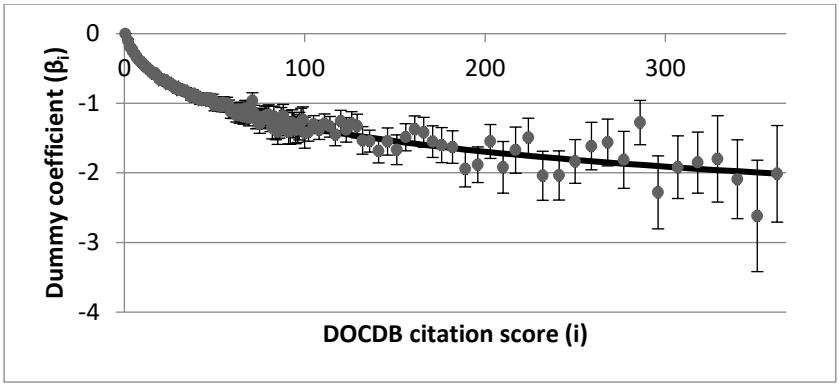

Figure 3: Estimates of the dummy coefficients related to different DOCDB citation scores, that were obtained from a Cox survival analysis, relating different scores on the DOCDB citation indicator to the maintenance time of a patent at the EPO. A 95% confidence interval is shown as well as a log-linear fit of $\beta_i = 1.27 - 0.55 \ln(10.80 + i)$, which has an R^2 of 0.93.

Similar results are found when repeating the analysis with EPO data (see Fig. 2). Here, the fit is even better with $R^2 = 0.93$ for the log-linear specification. However, the fits for the other curves also improve with $R^2 = 0.88$ for a quadratic curve and $R^2 = 0.78$ for a linear curve. Therefore, the found log-linear form appears to be robust with respect to the source of renewal data. Unfortunately, the analysis does not provide a guideline on the optimal offset, given that this parameter varies substantially with a value of 1.43 for the USPTO analysis and a value of 10.80 for the EPO analysis.

Robustness

In this paper, the functional form is estimated using Cox survival analyses. Unfortunately, these analyses rely on the proportional hazards assumption. This assumption states that the hazard function only differs by a constant non-time dependent value between the observed categories. This assumption can be verified by using Schoenfeld (1982) residuals and is violated severely (at $p<0.0001$) for both the USPTO and the EPO analyses.

In consequence, robustness tests using other model specifications have been performed. First, the binary approach from Hegde and Sampat (2009) is adopted, which determines the odds that a patent is renewed by a certain time. This approach has the benefit that it relies on few assumptions concerning the value function of the patent over time. Unfortunately, it also exploits less of the information contained in the renewal data by only observing the patent renewal time at one single time period. In this paper, a binary test is employed where the chance that a patent was renewed until it reached its maximum lifespan (20 years) is estimated using an adaption of Eq. (2), as shown below.

$$P(t_{renewal} = 20) = P\left(\varepsilon > C - \sum_i \beta_i \cdot DOCDB_i + \sum_{controls} \beta_{controls} \cdot x_{controls}\right)$$

If ε follows a logistic distribution, the b_i coefficients can be estimated using a logistic regression. Because this test relies on patents reaching their maximum lifetime, the sample is confined to patents that have the possibility of reaching it. At the USPTO, where full renewal is decided at 12 years, this includes patents with at least 13 years in the renewal data of spring 2014, which applies to all patents in the original sample. For EPO patents, the renewal decisions are taken yearly; hence, only patents with application years up to 1993 are included in the logistic analysis.

A linear regression analysis is also employed. Here, instead of estimating whether a patent is renewed at a certain point in time, its renewal time $t_{renewal}$ is directly estimated. This estimation can be constructed easily from Eq. (1), when assuming function $f()$ is linear, leading to the equation listed below.

$$t_{renewal} = C + \sum_i \beta_i \cdot DOCDB_i + \sum_{controls} \beta_{controls} \cdot x_{controls} + \varepsilon$$

A linear analysis is interesting because it facilitates use of the richness of the data - i.e. not just whether a patent is renewed but also for how many years - while still not having to rely on the proportional hazards assumption as is the case with the Cox survival analysis. Moreover, this analysis allows for direct computation of the size of the effect that patent citations have on the expected lifetime of a patent. Furthermore, both logistic and Cox survival analyses use a link function that relies on an exponential form. This could affect the form by which the citations correlate with patent renewal. Thus, a linear specification would be helpful in showing that the results found in ‘‘Results’’ section are not caused by modeling choices. Unfortunately, this linearization comes with the assumption that the value of a patent and its renewal time are linearly related, which is unlikely, as the ‘‘Methods and data’’ section explains.

This equation cannot be estimated using OLS because patents can only be renewed up to a certain point (i.e. 20 years). Therefore, the renewal time variable cannot take values greater than 20 years in our data, creating a need to deal with this censoring. Therefore, a Tobit regression analysis is employed, which considers censoring at the maximum lifetime of the patent, i.e. 20 years. Because this analysis has the same selection issues as logistic analysis, the sample is restricted in the same manner.

Finally, a lower bound of the value of a patent can be directly estimated using an interval regression analysis where the intervals are determined by the cumulative renewal fees, paid by the owner of the patent. Here again, censoring needs to be considered for patents that reach their maximum lifetime as well as censoring due to limited renewal data. Therefore, the sample restrictions for the binary analysis are also employed here.

The same analyses from the ‘‘Results’’ section are performed using both the Logit/Tobit and interval regression analyses rather than the Cox survival

regression. The same curves relating β_i to i are also estimated. The results from these estimations are presented in Table 3. From the evidence of these results, it is clear that the logarithmic functional form fits best the relation between the estimated coefficients of the dummy bi and the DOCDB citation score i. The analyses provide fit characteristics that are quite similar, indicating that the relation between patent citations and private value, as indicated by patent renewal, follows the same functional form regardless of the analysis. Therefore, it can be concluded that, of those studied, the log-linear form offers the best description of the functional form by which patent citations relate to patent value.

Table 4: R^2 of the different fits that relate β_i to i for each analysis at each patent office. It should be noted that the Logit, Tobit and interval analyses for EPO were performed on a smaller sample and are thus not fully comparable with the Cox survival regressions.

	USPTO				EPO			
	Cox	**Logit**	**Tobit**	**Interval**	**Cox**	**Logit**	**Tobit**	**Interval**
Linear	0.54	0.57	0.57	0.53	0.78	0.76	0.81	0.80
Qua-dratic	0.76	0.80	0.80	0.75	0.88	0.88	0.90	0.90
Log-linear	0.86	0.91	0.91	0.87	0.93	0.94	0.94	0.94

The relation between patent value and patent citations

The results obtained from the main analysis in the ‘‘Results’’ section, and from the robustness analyses, are informative in establishing the functional form that relates patent citations to patent value. However, the fits themselves may, in addition, explain the relevance of patent citations in more economic terms. Hence, the fits are presented with an interpretation of their estimated effect in Table 4.

The estimated effect of each additional patent citation is harder to estimate using the log-linear form. Therefore, the effects are given for patents that have double the number of patent citations than a patent with similar scores on the control variables. Given a log-linear fit of $\beta_i = a + b\ln(c + i)$ this translates as $b \ln 2$. It should be noted that having double the citations should be interpreted using the offset, i.e. the c parameter in the loglinear fit. Therefore, ‘doubling’ the citations for an uncited EPO patent means adding 14 citations in the case of the Tobit regression.

Table 5: Fits and economic interpretation of the analyses relating patent citations to patent value

Office	Analysis	Log-linear fit relating	Estimated comparative effect of having double the number of citations than a comparable patent
USPTO	Cox	$\beta_i = 0.33 - 0.32\ln(1.43 + i)$	Decreased abandonment hazard of 0.22
	Logit	$\beta_i = -0.43 + 0.48\ln(1.12 + i)$	Increased odds of full renewal of 0.33
	Tobit	$\beta_i = -3.15 + 2.86\ln(1.44 + i)$	Increased renewal time of 0.99 years
	Interval	$\beta_i = -1366 + 1642\ln(1.00 + i)$	Increased value of $1137.95
EPO	Cox	$\beta_i = 1.27 - 0.55\ln(10.80 + i)$	Decreased abandonment hazard of 0.38
	Logit	$\beta_i = -1.49 + 0.80\ln(6.49 + i)$	Increased full renewal odds of 0.56
	Tobit	$\beta_i = -8.64 + 3.32\ln(14.05 + i)$	Increased renewal time by 2.30 years
	Interval	$\beta_i = -12738 + 5021\ln(13.67 + i)$	Increased value of €3480.62

The results listed in Table 4 show that the estimated effect size of having been cited more than a comparable patent is substantial. Estimates show that patent citations confer value that can be measured in years of additional patent life and a value increase of thousands of euros/dollars. It should be noted that patent renewal analyses intrinsically estimate minimum values of patent value. Therefore, the value added by doubling patent citations may very well be much higher. Interestingly, with regard to USPTO patents, this value appears lower than for EPO patents, both in renewal time and patent value. However, the latter is in part due to the lower maintenance fees at the USPTO.

Applying the functional form to an econometric analysis

Introduction

The previous section established a log-linear relation between patent citations and patent value. Sets of patents, i.e. patent portfolios, can also be evaluated using patent citations. The value of a patent portfolio is generally estimated by counting the number of times any patent in the portfolio has been referenced. This practice could be justified on the assumption that the value of a patent portfolio is equal to the sum of the value of its members. The logical conclusion is that, when the value of individual patents is calculated differently, this should have repercussions for the estimation of the value of patent portfolios. Therefore, in this paper, a new method that relies on the found log-linear relation is introduced.

In this log-linear method, the value of a patent portfolio is derived by first computing a log-transformed value for each patent and then computing the sum of these log-transformed values. Because this method better models the relation between patent citations and patent value, it could prove superior to the normal linear method of estimating the value of a patent portfolio, i.e. simply computing the sum of individual patent citations.

A superior method of calculating the value of the patent portfolio would enhance understanding of firm innovative performance, an often-used metric in innovation research. To evaluate the log-linear method, a patent portfolio analysis is presented using both the traditional way of computing the value of a patent portfolio and the proposed log-linear method. For this endeavor, an adapted analysis of Hall et al. (2005) is presented, which relates Tobin's Q, (Q) to stocks of R&D, patents and patent citations.

Methods and data

The analysis adapted from Hall et al. (2005) will be used to assess which method of evaluating patent portfolios better explains firm performance: the common linear method, or the log-linear method that models a log-linear relation between patent citations and patent value. The analysis of Hall et al. (2005) models Tobin's Q as a function of the relative knowledge stock of the firm. This stock is then approximated using the ratio between the R&D stock and the assets of the firm as well as other ratios involving the R&D stock, the patent stock, and the patent citation stock. Controls for year, as well as the firm, will also be included in the analysis. Therefore, the following equation is estimated:

$$\ln Q_{it} = C_i + C_t + \ln\left(1 + \beta_1 \frac{RandD_{it}}{A_{it}} + \beta_2 \frac{PAT_{it}}{RandD_{it}} + \beta_3 \frac{CITES_{it}}{PAT_{it}} + D_{RandD_{it}=0}\right) + \varepsilon_{it}$$

Here C_i and C_t denote constants of firm i and time t while A_{it} denotes the total assets. $RandD_{it}$, PAT_{it} and $CITES_{it}$ denote respectively the R&D, patent and citations depreciated stock. $D_{RandD_{it}=0}$ denotes a dummy for firms with no reported R&D expenditures at time t. When this dummy is equal to 1, the ratio $\frac{PAT_{it}}{RandD_{it}}$ is set to 0 if the R&D stock, $RandD_{it}$, is equal to 0. There are also cases for which the patent stock is equal to 0; these are not used in the analysis. Finally, ε_{it} represents a random error.

The non-linear analysis that follows from equation 4 is presented along with a linearized version, which assumes $\ln(1 + x) \approx x$. This linear analysis has the benefit of facilitating a fixed effects approach, which is not possible with the non-linear analysis, as noted in Hall et al. (2005). In the non-linear analyses, C_i is approximated using sector dummies of the Standard Industry Classification (SIC).

For this analysis, USPTO applications (since the sample mainly concerns US firms), combined with DOCDB citation information are matched to a random sample of patenting firms with at least 100 patents listed in PATSTAT, and that are listed in the Compustat database. For the resulting sample of 1092 firms, financial data is considered from the years between 1981 and 2005. In this paper, citation stock increases are modeled using a linear model, as well as the log-linear model previously specified.

The sample was constructed as follows: only firms that have a continuous presence in at least two periods in the dataset were used. Moreover, in order to accurately compare patenting firms, only observations of firms that have a non-zero patent stock - i.e. observations of firms that have at least one patent in the current or any previous period - are used.[22] Finally, Tobin's Q was not known for all observations in the resulting dataset, leading to the removal of 1890 observations.

The R&D stock was initialized as the R&D expenditure for the first year in which a firm enters the sample divided by 0.23, in a procedure similar to Hall (1990), and Hall et al. (2005). The patent stock and the citation stocks are not initialized because the full patenting activity of all firms is observed for 30 years prior to the first year of the sample using the EPO PATSTAT database. Finally, all stocks are depreciated by 15% each year, in line with Hall et al. (2005). The descriptive statistics of the sample are detailed in Table 5.

[22] A comparative analysis that included observations for firms with no patent stock as well as a dummy controlling for this occurrence yielded very similar results to the analyses presented in this paper.

Table 6: Descriptive statistics of the variables used in the horse-race regressions. Below each citation stock is listed the formula used to create it, where CIT refers to the citation score of an individual patent. R&D stock and total assets are adjusted for inflation using USBLS(2016) data (1983=100). All stocks are calculated with a 15% depreciation rate.

Variable	Description	Number of Observations	Mean	Standard Deviation	Min	Max
Ln (Tobin's Q)	Natural log of market value divided by total assets	13044	0.39	0.78	-6.23	4.63
Year	Book year of the firm, application year of the patents	13044	1995	7.39	1980	2005
R&D stock	The current stock of R&D expenses ($M)	13044	631	2091	0.00	29814
Total Assets	The total assets of the firm ($M)	13044	4841	24430	0.04	658800
D(R&D=0)	Dummy to indicate no R&D expenses in that year	13044	0.116	0.32	0	1
Patent stock	The current stock of USPTO patents	13044	239	843	0.00	14649
Citation stock $\sum CIT$	The current citation stock, calculated using the linear method	13044	4402	15156	0.02	226776
Citation stock $\sum \ln(1 + CIT)$	The current citation stock, calculated using the log-linear method	13044	579	2038	0.01	31932

The results of the fixed effects linear models are shown in Table 6. The increase in R^2 shows that citation indicators using the log-linear transformation perform better - with an increase of 3.4% in explained variance - at explaining log Tobin's Q than the citation indicators without the use of the logarithmic transformation. This represents a substantial increase of 70% in added explained variance by introducing a citation indicator to explain company performance. Therefore, this analysis shows the potential of applying the log transformation to portfolio analysis, while simultaneously providing external validity to the findings in the ''Measuring the relation between citations and renewal'' section.

The non-linear analysis of Hall et al. (2005) was also performed: see analyses 4, 5 and 6. Applying their analysis to this paper's sample produces very similar results, with one exception: the ratio of patent stock over R&D stock, representing patenting efficiency, is negative. The likely cause is the inclusion of several firms for which R&D expenditure is not listed in the COMPUSTAT database, and for which this ratio is recorded as 0. Analyses excluding these firms produce a positive coefficient for this ratio. The linear and the loglinear specifications of the citation stock perform very similarly in the non-linear analysis: there is only a difference of 0.007 in their R^2. Using the log-linear form only adds 2.9% in added explained variance.

Including firm dummies, instead of SIC dummies, give very similar results, as can be seen from analyses 7, 8 and 9. However, here the linear specification performs slightly (0.006) better. This represents a decrease of 17% in explained variance when using the loglinear method as opposed to the classical method. This analysis, therefore, demonstrates that that the log-linear method of citation counting does not always deliver improvement, but it produces adequate results nonetheless.

Table 6: Horse-race regressions explaining Ln(Tobin's Q). Variables with [a] represent stocks with a 15% depreciation rate. Below each citation stock is listed the formula used to create it, where CIT refers to the citation score of an individual patent. Cluster-robust standard errors are reported in parentheses and asterisks indicate statistical significance with: * p < 0.05, ** p < 0.01, *** p < 0.001.

	(1) Fixed effects	(2) Fixed effects	(3) Non-linear	(4) Non-linear	(5) Non-linear	(6) Non-linear
R&D[a]/ Total assets	0.0156**	0.0156**	0.145**	0.237*	0.0439	0.0558
	(0.00550)	(0.00546)	(0.0463)	(0.0922)	(0.0316)	(0.0406)
Patents[a]/R&D[a]	-0.000743	-0.000781*	-0.00055**	-0.0011***	-0.000544**	-0.000723***
	(0.000387)	(0.000324)	(0.000181)	(0.000222)	(0.000192)	(0.000180)
D(R&D$_{it}$=0)	0.0811	0.0767	0.0422	0.0620	0.0960	0.127
	(0.0666)	(0.0652)	(0.0588)	(0.119)	(0.0825)	(0.105)
Citations[a]/Patents[a] $\sum$ **CIT**	0.00213***		0.0119***		0.00457**	
	(0.000579)		(0.00181)		(0.00155)	
Citations[a]/Patents[a] $\sum \ln(1 + CIT)$		0.124***		0.630***		0.159*
		(0.0308)		(0.152)		(0.0642)
Constant	0.204***	-0.0601	0.0954	-0.540**	-0.355***	-0.567***
	(0.0333)	(0.0834)	(0.122)	(0.201)	(0.0177)	(0.0877)
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Firm controls	Yes	Yes	No	No	Yes	Yes
SIC dummies	No	No	Yes	Yes	No	No
N	13044	13044	13044	13044	13044	13044
Nr. Firms	1092	1092	1092	1092	1092	1092
Nr. SIC	214	214	214	214	214	214
R^2	0.125	0.160	0.4905	0.4912	0.7052	0.7046

Conclusion

The main result of this paper is that patent citations display a log-linear relation with patent value. Therefore, researchers are advised to take this relation into consideration when using patent citations to approximate patent value. The fits obtained from the renewal analysis show that patents with double the number of citations of comparable patents, have an increased value of $1137.95 in the case of USPTO patents and €3480.62 in the case of EPO patents.

The results of the firm analysis indicate that, at least in some economic models, it may be better to first apply a log transformation to the citation count of an individual patent before computing the sum. Doing so may yield an improvement of up to 70% in added explained variance. Yet, in another analysis, the classical way of calculating patent citations has proved slightly superior. For that reason, the log-linear transformation should be used with caution.

When using a logarithmic functional form, the explanatory power of the citation indicator improves. Yet, much unexplained variance remains. Therefore, we should continue to keep in mind the limited ability of patent citations to approximate patent value. Moreover, the found functional form reflects the relation between patent citations and private value, but it may not hold true for other value constructs such as the social value and (knowledge) impact of a patent. Hence, researchers should be careful when applying the findings of this paper to approximate other constructs of patent value.

References

Albert, M. B., Avery, D., Narin, F., and McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20(3), 251–259.

Albrecht, M. A., Bosma, R., van Dinter, T., Ernst, J. L., van Ginkel, K., and Versloot-Spoelstra, F. (2010). Quality assurance in the EPO patent information resource. *World Patent Information*, 32(4), 279–286.

Arts, S., Appio, F. P., and Van Looy, B. (2013). Inventions shaping technological trajectories: Do existing patent indicators provide a comprehensive picture? *Scientometrics*, 97(2), 397–419.

Bakker, J., Verhoeven, D., Zhang, L., and Van Looy, B. (2016). Patent citation indicators: One size fits all? Scientometrics, 106(1), 187–211.

Barabási, A. L., and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439),509–512.

Belenzon, S. (2012). Cumulative innovation and market value: Evidence from patent citations. *The Economic Journal*, 122(559), 265–285.

Bessen, J. (2008). The value of US patents by owner and patent characteristics. *Research Policy*, 37(5), 932–945.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30, 89–99.

Carpenter, M. P., Narin, F., and Woolf, P. (1981). Citation rates to technologically important patents. *World Patent Information*, 3(4), 160–163.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359), 557–565.

Gambardella, A., Harhoff, D., and Verspagen, B. (2008). The value of European patents. *European Management Review*, 5(2), 69–84.

Gay, C., and Le Bas, C. (2005). Uses without too many abuses of patent citations or the simple economics of patent citations as a measure of value and flows of knowledge. *Economics of Innovation and New Technology*, 14(5), 333–338.

Gittelman, M. (2008). A note on the value of patents as indicators of innovation: Implications for management research. *The Academy of Management Perspectives*, 22(3), 21–27.

Hall, B. H. (1990). The manufacturing sector master file: 1959–1987 (No. w3366). *National Bureau of Economic Research*.

Hall, B. H., Jaffe, A., and Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics*, 36, 16–38.

Harhoff, D., Narin, F., Scherer, F. M., and Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, 81(3), 511–515.

Harhoff, D., Scherer, F. M., and Vopel, K. (2003). Citations, family size, opposition and the value of patente rights. *Research Policy*, 32(8), 1343–1363.

Hegde, D., and Sampat, B. (2009). Examiner citations, applicant citations, and the private value of patents. *Economics Letters*, 105(3), 287–289.

Hertz-Picciotto, I., and Rockhill, B. (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, 53, 1151–1156.

Hsieh, F. Y. (1995). A cautionary note on the analysis of extreme data with Cox regression. *The American Statistician*, 49(2), 226–228.

Hung, S. W., and Wang, A. P. (2010). Examining the small world phenomenon in the patent citation network: A case study of the radio frequency identification (RFID) network. *Scientometrics*, 82(1), 121–134.

Jaffe, A. B., and de Rassenfosse, G. (2016). Patent citation data in social science research: Overview and best practices (No. w21868). *National Bureau of Economic Research*.

Lanjouw, J. O., Pakes, A., and Putnam, J. (1998). How to count patents and value intellectual property: The uses of patent renewal and application data. *The Journal of Industrial Economics*, 46(4), 405–432.

Magerman, T., Van Looy, B., and Song, X. (2006). Data production methods for harmonized patent indicators: Patentee Name Harmonization. *EUROSTAT Working Paper and Studies*, Luxembourg.

Maurseth, P. B. (2005). Lovely but dangerous: The impact of patent citations on patent renewal. *Economics of Innovation and New Technology*, 14(5), 351–374.

Neuhäusler, P., and Frietsch, R. (2012). Patent families as macro level patent value indicators: Applying weights to account for market differences. *Scientometrics*, 96(1), 1–23.

Pakes, A. (1986). Patents as options: Some estimates of the value of holding European patent stocks. *Econometrica*, 54(4), 755.

Pakes, A., and Schankerman, M. (1984). The rate of obsolescence of patents, research gestation lags, and the private rate of return to research resources. In Z. Griliches (Ed.), *R&D, patents, and productivity* (pp.73–88). University of Chicago Press.

Peeters, B., Song, X., Callaert, J., Grouwels, J., and Van Looy, B. (2010). Harmonizing harmonized patentee names: an exploratory assessment of top patentees. *Eurostat Working Paper*.

de Rassenfosse, G., and Jaffe, A. B. (2014). Are patent fees effective at weeding out low-quality patents? *(No.w20785)*. National Bureau of Economic Research.

de Rassenfosse, G., and Van Pottelsberghe de la Potterie, B. (2013). The role of fees in patent systems: Theory and evidence. *Journal of Economic Surveys*, 27(4), 696–716.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1), 239–241.

Thomas, P. (1999). The effect of technological impact upon patent renewal decisions. *Technology Analysis and Strategic Management*, 11(2), 181–197.

Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The RAND Journal of Economics*, 21(1), 172–187.

United States Bureau of Labor Statistics. (2016). Consumer price index–All Urban Consumers, not seasonally adjusted, series id: CUUR0000SA0. http://data.bls.gov/timeseries/CUUR0000SA0. Accessed 1 September 2016.

Van Pottelsberghe de la Potterie, B., and Van Zeebroeck, N. (2008). A brief history of space and time: The scope-year index as a patent value indicator based on families and renewals. *Scientometrics*, 75(2), 319–338.

Van Zeebroeck, N. (2011). The puzzle of patent value indicators. *Economics of Innovation and New Technology*, 20(1), 33–62.

Patent citations and a framework of the productive and market value of patents

Jurriën Bakker and Bart Van Looy

Abstract

We advance a novel framework in which patent value is composed of a productive component, i.e. the benefits stemming from creating and selling goods based on patented inventions, and a market component, i.e. the direct or indirect benefits accrued to the owner of a patent by means of licensing, litigating or trading. Building on the legal role of patent citations, we propose that productive value can be approximated by self-citations, while market value can be estimated by the citations a patent receives from other patents (non-self-citations). We then derive hypotheses about the likely actions (renewal, sales and licenses) of patent owners, dependent on the productive and market value of their patents. We procede to validate our framework by testing these hypotheses on a large body of EPO (N= 547,365) and USPTO patents (N=571,816), filed between 1981 and 2000. The obtained results support our assertions. The relevance of the advanced framework is then shown by applying it to the strategy of patent owners during the life of their patent.

Keywords: Patent citations, Patent value, Patent renewal, IP strategy, Patent Licenses, Patent transfers

JEL Classification O34

Acknowledgements: We would like to thank Diana Ferreira, Otto Toivanen, Dirk Czarnitski and Koen Frenken for their comments on this paper. Additionally, we would like to thank Adrián Kovacs for providing data on the transfers of USPTO patents.

Introduction

Measuring innovative performance is of crucial importance in our current knowledge driven economy. Most researchers opt to use patent statistics for this endeavor, as they are a vast and detailed source of information on innovation. However, not all patents have the same relevance, which forces researchers to determine their value to arrive at a proper measure of innovation (Griliches,1990). Determining the value of patents is a difficult endeavor, given that they are intellectual property. Patents are related to a vast number of different technologies and they can be owned by firms, non-profits and individuals. Additionally, it is often difficult to determine their exact share in the profits of a product or firm. Moreover, for researchers and managers of large portfolios, it is necessary that large sets of patents can be valued quickly and consistently.

In this paper, we will define the value of a patent as the added utility that an owner derives from the possession of the patent. It is to be noted that this definition, better known as the patents private value, excludes other value constructs such as social value, i.e. the value to society, and inventive value, i.e. the novelty of the technological process protected by the patent (see e.g. Carpenter et al., 1981; Fleming, 2007; Verhoeven et al., 2016). Patent owners can generally[1] derive value from their patents in two ways: first, they can use the patent to deter competition from their products and thus gain additional profits; second, patents can be monetized through engaging in transactions, licenses, or by suing those that infringe. We will refer to these two value constructs respectively as the productive value and market value of a patent.

We argue that these constructs of patent value can be measured by observing the amount of times a patent has been referenced by other patents. The forward citations indicator is a viable candidate to observe our value constructs because it has been clearly established in the literature that it is a consistent indicator of the value of a patent (Carpenter et al., 1981; Trajtenberg, 1991; Jaffe et al., 1992; Jaffe and Trajtenberg, 1999; Harhoff et al., 1999; Thomas, 1999; Gambardella et al., 2008). Likewise, this relationship has been established for several constructs of patent value: inventive value (e.g. Carpenter et al., 1981), and social value (Trajtenberg, 1991), as well as the general private value of patents (e.g. Harhoff et al., 1999; Thomas, 1999; Hall et al., 2005; Gambardella et al., 2008), and finally the sales price of the patent (Fischer and Leidinger, 2014).

Initial contributions (e.g. Trajtenberg, 1990; Jaffe et al., 1992; Jaffe and Trajtenberg, 1999) argued that patent-to-patent citations reflect spillovers: ‘‘citing patents would bear a sort of causal relationship to the cited patent, with citations being the overt manifestation of such a link’’ (Trajtenberg, 1990 p.

[1] Patents can also be used as collateral (Amable et al., 2010) and even for reduced tax loads (Karkinsky and Riedel, 2012).

185). Jaffe et al. (1993 p. 578) argue that ‘‘knowledge flows do sometimes leave a paper trail, in the form of citations in patents’’. Thus, a citation between two patents means that the cited patent represents previously existing knowledge upon which the citing patent builds (see also Jaffe et al., 2000). Conceived as such, highly cited patents reflect impact, relevance and hence quality, similar to the role citations play in the scientific literature[2].

However, unlike scientific citations, patent citations (front page references) are the result of a search and selection process of prior art by the examiner, not by the inventor, and ultimately serve a legal purpose (Callaert, Pellens and Van Looy, 2014, but see also Michel and Bettels, 2001; Tijssen et al., 2000; Meyer, 2000). References in the scientific literature are added by the author during their research, and serve to indicate and acknowledge previous knowledge on which the contribution builds. References in patents are added during the granting process for evaluating novelty and inventiveness, and for qualifying the claims made in the patent. The cited prior art on the front page of patent documents is ultimately selected by the patent examiner and not by the applicant/inventor, who might, or even should (in the case of USPTO applications, the so called ‘duty of candor’), bring relevant references to the attention of examiners. Based on available information in archives and databases, patent examiners ultimately decide which references are relevant to include for qualifying the claims implied in the patent document. Therefore, it would be unwise to consider patent citations as a direct analogy to academic citations.

It would be better to evaluate patent citations for what they are: references that operate within the legal context of the patent application process. Patent citations will often signal that the technological scope of the citing patent is reduced by the cited patent. Even if the scope of the citing patent remains intact, patent citations are still an indication that the citing and the cited patent encompass substantially related technology. How these notions can be used to link the forward patent citations to the productive and market value of the patent will be explained in the following paragraphs.

[2] Albeit that both recall and precision deserve attention: Jaffe et al. (1993) already signal issues when conceiving patent citations as knowledge spillovers- not all spillovers are captured in citations and not all citations represent direct knowledge spillovers, as the more recent study of Callaert, Pellens and Van Looy (2014) reveals (see also Alcacer and Gittelman, 2006; Breschi and Lissoni, 2001 in this respect).

It is likely that when an owner greatly values the products that the patent protects, they will spend additional resources in expanding the technology. This additional effort can then be observed by self-citations, i.e. the number of times the patent is cited by other patents belonging to the same owner. Supporting this theory, (Narin et al., 1987) found that these citations are often associated with patents that are used in continuous and ongoing projects. In addition, Belenzon (2012) found that these self-citations are indicators of a cumulative effort within the firm, with Thomas (1999) finding that self-citations may explain patent value better than non-self-citations. Therefore, self-citations are a promising indicator to signal the *productive* value of a patent.

We propose that non-self-citations on the other hand, are an indication of *market* value because they indicate the position of the patent document within the legal context of the patent system. In a similar vein to the work on patent thickets by von Gravenitz et al. (2011), we propose that non-self patent citations indicate value because they limit the scope of citing patents. Since many technologies are composed of multiple patentable inventions, a patent limiting the scope of subsequent patents can be very valuable. This is because it can prevent or hinder the production of products which are based on the citing patents, if they still partly rely on the claims defined in the cited patent. This blocking power of the cited patent can then be translated in demands for compensation from the owner of the blocked product. Therefore, a patent being cited indicates that its owner may draw value from others and thus these patent citations should indicate market value.

In the remainder of the paper we will first relate our value framework to the decisions of patent owners. Then we will expand on the measurement of patent value using patent citations, to arrive at four testable hypotheses. These hypotheses will then be tested using a large set of data on EPO and USPTO patents. Finally, we will show the power of the framework in explaining when owners are interested in using patents to produce or gain monetization from the market.

Theory and hypotheses

In this section, we will link productive and market value to the major decisions patent owners face over the lifetime of the patent. Next, we will link patent citations to productive and market value. Combining these exercises, we will derive testable hypotheses that we can use to validate our framework.

Patent value and owner operations

Patent renewal

Most patent offices, notably the EPO and the USPTO, charge maintenance fees for patents to remain in force until their maximum time. It is expected that owners will maintain their patents for as long as they believe that keeping the patent outweighs the costs of maintaining it. Patent renewal has long been recognized as a primary indicator of the private value of a patent (e.g. Pakes, 1986; Thomas, 1999; Harhoff et al., 1999; Mauseth, 2005; Hegde and Sampat, 2008). In our framework, the private value of a patent is composed of its productive value and its market value. Therefore, we expect that patents that have a high productive and/or market value will be maintained, while patents that have neither will be abandoned.

Patent transfers

Patents, being a defined property, can be transferred from one owner to the other. In doing so, the original owner relinquishes all rights to his invention and in exchange obtains a lump sum payment. The rationale for selling can be easily understood when an owner has more to gain from selling the patent than from using it to produce goods (Serrano, 2010). This will generally happen if the buying party has a comparative advantage (e.g. in the marketing or production capabilities) in the technology protected by the patent (Gans and Stern, 2000; Gans et al., 2002; Galasso et al., 2013; Figueroa and Serrano, 2013). This immediately reveals that productive and market value impact the decision to sell differently: a higher market value increases the chances that the patent is sold, due to the higher possible returns for the owner. A higher productive value on the other hand, represents higher opportunity costs and is therefore detrimental to the chances a patent is sold.

However, there is more to patent transfers than it appears from this simple analysis. For an owner to transfer a patent, they need to consider all possible revenue for the patent into the future (Smith and Parrs, 2005). Moreover, the owner needs to consider that selling the patent may inhibit them from building further on the technology, since they will be forced to deal with the patent's buyer for its use (Heald, 2005). Therefore, it is unlikely that an owner will sell any patent, or patent portfolio, if they believe that in some future they will want to expand on its technology. Thus, patents with a high productive value are unlikely to be sold.

Patent licenses

Besides selling a patent, and thereby completely relinquishing control, owners may also choose to license their patents. This entails allowing a party to use technology defined in the patent, under certain conditions, in exchange for a monetary compensation (e.g. a royalty or a fixed fee), or even a non-monetary compensation in the case of cross-licensing.

Patent owners will decide to license their patent based on the invention and the size of the market. According to Katz and Shapiro (1985), as well as Rockett (1990), licensed patents should be of lesser quality since the creation of a competitor will eat into their monopolistic profits, which is especially true for larger firms (Arora and Fosfuri, 2003). On the other hand, licenses should be interesting for smaller firms as they may not have the productive capabilities to commercialize the invention (Sakakibara, 2010).

The decision to license a patent is also contingent to the position of the patent owner in the market: if the owner is an outsider in the market where his patent upholds, thus not a competitor, licensing can be done in a similar method as a sale, i.e. with a one-off payment (Kamien and Tauman, 1986; Katz and Shapiro, 1986; Kamien et. al., 1992; Kamien, 1992); if on the other hand the owner is an insider in the market, he will become a natural competitor. In that case it is however, still possible for the owner to profit though licensing by using a royalty structure (Rockett, 1990; Wang, 1998; Poddar and Sinha, 2001; Kamien and Tauman, 2002). However, the theoretical derived license structures do not always hold up in practice, as found by Rostoker (1983), and Taylor and Silberston (1973).

It is obvious that market value is of interest for the chances of licensing, since owners will only license if there is sufficient interest, albeit that this relation may be moderated by owner specific variables. The productive value could be negatively related to licenses, since many authors refer to the trade-off between licensing revenues and the profits from increased market power. But, unlike patent transfers, owners may still license their patents even if they have established a position in the market, when royalty payments outweigh their losses in market power.

A framework of patent value and patent decisions

It is possible to combine the insights of the previous subsection into one framework. This is done by reviewing the likely decision of a patent owner as a function of the productive value and the market value of a patent. These considerations of the owner can be graphically represented, as shown in table 1.

Table 7: Conceptual framework of the decisions of patent owners in relation to the productive and market value of their patent.

Productive value	Market value: *Low*	Market value: *High*
High	**Maintain**	**License**
Low	**Abandon**	**Sell**

Patent citations and their relation to distinct kinds of patent value

Non-self-citations as an indicator of market value

As stated in the introduction, patent citations serve a legal purpose as they are essential to interpret the claims of the citing patent. Interestingly, this concept is rarely utilized in the literature in lieu of the interpretation of patent citations as knowledge sources. To better understand the (legal) role of patent citations, we present an overview of non-self-patent citations and their relation to patent value.

The literature regarding the processes that lead to patent citations focusses on citations that are either given by the applicant or the examiner. Therefore, we will also use this distinction in our overview. However, it bears reminding that the examiner has the final judgement on which citations end up on the 'frontpage', and thus in many patent databases. Given that applicants are the first to deliver their references, if any, we will start with their reasons for doing so.

Applicants may cite patents to obtain an expert judgement on the validity of their patent considering these cited patents (Akers, 2000). Interestingly, applicant submitted prior art is also often ignored by patent examiners

(Cotropia et al., 2013), which implies that the expert judgement is often not fully provided. Therefore, applicant citations may imply serious opportunities for the holders of the cited patent to extract rents for possible infringement by the holders of the citing patent. This signal of litigious opportunity, and thus potential value may be further enhanced, since Lampe (2012) and Sampat (2013) show that applicants will cite more frequently when their patent has a high value.

However, applicants may be strategic in the citations they provide (Lampe, 2012), as they may be unwilling to provide relevant prior art (especially of competitors), if this prior art would be instrumental in rejecting or narrowing down their patent (Sampat, 2005). Finally, applicants may also try to 'flood' the patent office with citations of inferior quality, i.e. less relevant to the patentability of their patent, to limit the time an examiner spends on finding possibly limiting prior art. A theory that is partially confirmed by Cotropia et al. (2013).

Examiners on the other hand, do not appear to rely greatly on the citations of applicants and perform in any case an independent search for prior art (Alcacer and Gittelman, 2006; Azagra-caro et al., 2011; Cotropia et al., 2013). Instead they cite according to the region where the patent is from, the volatility[3] of its technology class (Tan and Roberts, 2010), and the economic sector of the patent, albeit that they may still be influenced by the included references of the applicant (Azagra-caro et al., 2011). When examiners cite a patent, this may be viewed as a more objective measure of the relevance of the patent, since examiners are tasked explicitly with using prior art as a legal basis to reject or narrow patents (e.g. Schmoch, 1993). In addition, they may be further apart from the current technological frontier than the applicant (Eisenberg, 2004) and live in a different legal reality (Noveck, 2006).

Nonetheless, there may also be noise in examiner citations: Cockburn et al. (2003) found that there are substantial differences in citing behavior between examiners, with Lemley and Sampat (2012) explaining the examiners' tenure as one of sources of these differences. Additionally, examiner citations also provide avenues for the citing patent to work around the cited patent without infringing it (Tan and Roberts, 2010). Finally, examiners may also have 'favorite' patents they often cite which may not always be the most relevant to the citing patent (Cockburn et al., 2003)

From this brief overview, we conclude that the most likely reason most patent citations are introduced is because the cited patent is relevant in some way to the citing patent, albeit that this relation may be moderated by the familiarity of the cited patent in the community as well as the value of the citing patent. Notwithstanding, as we also discussed, certain practices introduce a large degree of noise in this relation, both for examiner and applicant generated citations. This makes it difficult to determine whether applicant or examiner

[3] Tan and Roberts measure the volatility of a technology class by the amount of revisions that are made to it by the patent office.

citations are better at indicating market value. In any case, the examiner ultimately decides on the included citations, primarily for legal purposes. Therefore, we will not introduce a difference between the citations in our framework.

The notion that patents which are relevant to other patents are likely to have a market value, follows from the fact that since the cited patent is in principle filed before the citing patent, it may lay a claim on the knowledge space of the citing patent. If the owner of the citing patent uses this part of the claimed knowledge space of his patent then the owner of the cited patent may claim damages or demand license or acquisition of the cited, and now infringed, patent. Therefore, a patent being cited is likely to be correlated with the possibility to monetize the patent, and therefore the market value of the patent.

Self-citations as an indication of productive value

Patents can also be cited by patents that belong to the same owner, this is referred to as a self-citation. Naturally, this does not involve market value since owners are not going to demand compensation from themselves[4].

Since early patent citation literature (Narin et al., 1987), it has been argued that self-citations provide a valuable indicator of innovative quality. A patent with many self-citations indicates that the owner pursued further activity in the area. It is thus likely that patents with many self-citations indicate building stones for large projects within the firm (Narin et al., 1987; Lanjouw and Schankerman, 2004). Moreover, it has also already been shown, that a count of self-citations performs well at explaining patent value (Bessen, 2008) and may even outperform a count of other citations (Thomas, 1999; Hall et al., 2005; Belenzon, 2012).

In conclusion, the literature mostly suggests that a patent that receives many self-citations will be of a more productive nature, as it then constitutes a part of a larger patent portfolio of the owner of the patent. It is thus likely that the patents that are cited multiple times by patents from the same owner protect a critical part of the innovative effort of the owner and therefore possess a substantial productive value.

Hypotheses

From the discussed theories relating patent value to decisions of patent owners, and patent citations to patent value, we can derive several hypotheses that can be used to validate our framework. These hypotheses can then be tested by analyzing patent data. In this subsection, we will briefly discuss each of the hypotheses and their underlying rationale.

Productive value and market value as different constructs

Throughout the theoretical section, we assumed that productive value and market value classify as substantially different constructs of patent value. It is

[4] Internal transfers do happen, for instance because of fiscal considerations.

likely that these constructs have a degree of correlation. After all, if a patent has a substantial value to its owner, it is likely that other parties are also interested in acquiring or licensing it. Nonetheless, for our framework to be useful, the constructs must be substantially different from each other. Therefore, in the data we expect that our measures, self-citations and non-self-citations, contain different information, i.e. one indicator does not explain a large amount (>50%) of the variation in the other.

Patent transfers and patent licenses are positively correlated by market value

Patent transfers and patent licenses can only occur by the grace of sufficient interest from others. Therefore, market value should positively impact the chances of a patent being sold or licensed, and thus non-self-citations should correlate positively with patent transfers and patent licenses.

Patent transfers are negatively correlated with productive value

Productive value is unlikely to positively relate to the sale of a patent because productive value, in this case, represents the opportunity costs for the patent owner: when they sell the patent they are unlikely to be able to use its technology and obtain any revenue from products related to it. Therefore, the productive value of a patent, as measured by its self-citations, should be negatively correlated with its chances of being sold.

Productive and market value correlate with patent renewal

Patent owners are faced with the decision of continuing to pay maintenance fees for the patent, which they will continue to do if the value of the patent outweighs the costs of maintaining it. In our framework, productive and market value should both increase the value of the patent for its owner. Therefore, patents that have a high productive and/or market value should be more likely to be renewed. We hence expect that both citation based measures of patent value, self-citations and other citations, correlate positively with patent renewal.

Methods and Data

Data and sample selection

We use patent citation indicators constructed from the October 2013 version of the EPO PATSTAT database. We want to perform and compare analyses at both the EPO and the USPTO. Consequently, the sample was constructed of DOCDB patent families that contain at least an EPO application and at least one USPTO application. This allows for a sample in both offices of identical inventions, since all members of this family have the same technical content (Albrecht et al., 2011). It is to be noted that this decision biases the dataset to include more valuable patents, as patents with larger families have been found to be of higher value (e.g. Harhoff et al., 2003). We further restricted the families to have granted patents in both EPO and USPTO, since patents are likely to be abandoned at some point if they are never granted. We then applied a basic data cleaning (equal to Bakker et al., 2016) to remove duplicates caused by untraceable priorities and citations, incorrect conversions of patent numbers, and several issues caused by changes in the USPTO system in 2001. The changes brought by this data-cleaning are minor but ensure that all patents and citations in our dataset correspond to actual patents filed in the relevant patent offices.

We decided to have EPO patents valued by the citations in the EPO system, and the USPTO patents by the citations in the USPTO system. This is done to ensure the patent is valued in its legal context, rather than as a general measure of quality. Bakker et al. (2016) show that this choice represents a substantial difference in the patent citation counts, and therefore our results should always be considered as representing the value of patents in their own patent system.

To enable a decent time window for observing renewal decisions, sales, licenses and patent citations, we only included patents up until the year 2000. The sample was further restricted to include only patents that were applied later than 1981 because we are unsure of the accuracy of renewal data of earlier patents[5]. We therefore arrive at a dataset with patents between 1981 and 2000. It is to be noted that even after these procedures, a small number of patent families had more than one member at the same patent offices. Therefore, the number of observations between analyses using EPO indicators (N=547,365), is slightly different than those that rely on USPTO indicators (N=571,816).

Licenses and sales data

In this paper, we explain patent licensing deals and patent sales using a model of market and private value. We use data from the INPADOC PRS file of the spring version of PATSTAT 2014. Unfortunately, from this data we could only extract licenses and sales of EPO patents. For this reason, we added license

[5] We observed a larger number than expected of never renewed patents for earlier dates. This indicates that renewal data may be incomplete, thus making an analysis of these years unreliable.

and transaction data from the research of Marco et al. (2015) which relate to USPTO documents.

A major problem with the analysis of patent sales is that, even though it is sometimes obligatory to license changes in ownership to the patent office, this is often omitted by the parties involved (Kovács, 2016). This entails that we need to treat the reported data of sales and licenses as inherently incomplete and according to Kovács (2016) biased towards patents of higher value. Therefore, we should not focus on whether higher valued patents are sold, but instead if productive and market value relate differently to the chances of a patent being sold.

It is likely that registrations of licenses are even more incomplete, as no requirement generally exists to register licenses of patents. Here we will instead analyze if the chance of licenses is explained by both an increase in market value and an increase in private value.

Methods for estimating the effect of patent citations on patent renewal

To observe patent renewal, we will use data on maintenance payments recorded in the INPADOC PRS file, provided by the 2014 April version of the EPO PATSTAT database. Patent renewal as an indicator is derived from the process in which the owners of a patent need to pay periodical fees to keep a patent in force. For patents at European offices this fee is paid yearly (after an initial free period of a few years), while owners of USPTO patents only pay this fee at 4,8 and 12 years after application. The value of patent documents can then be assessed by observing if any renewal fees have been paid and for what period they were paid. It is to be noted that patent renewal is (in general) limited by the maximum lifetime of patents - 20 years. This entails that the patent will expire, regardless of the desire of an owner to pay any price for renewal at this point. The implication of this expiration is that all patents, which have a value/revenue stream above the threshold to be maintained until their twentieth year, will have an exactly equal score on the renewal indicator. Therefore, a censoring occurs with respect to the higher values that may be attributed to the patent.

In this paper, we will describe renewal as the end of the period for which maintenance fees have been registered. We observed patent renewal for all granted patents of both EPO and USPTO by examining renewal fees paid to the respective office[6]. In the case of EPO patents, we considered a patent renewed if it has been renewed in at least one national office which subscribes to the EPO regional system. This method follows the Single Renewal Approach (SRA), as discussed in van Zeebroeck (2011) and which was also used in Bakker (2017).

[6] i.e. we observed the last registered renewal payment of the patent at its respective office.

We tested the robustness of the SRA method in appendix C by using instead data of a single patent office (i.e. the German, British, and French patent office). In this analysis, we found that the results obtained by only using patents from one of the European patent offices does not differ substantially from those presented by aggregating the offices using the SRA method.

The patent renewal indicator is often seen as an indicator of value (e.g. Pakes and Schankerman, 1984; Pakes, 1984; Lanjouw et al., 1998; Harhoff et al., 1999; Thomas, 1999; Hegde and Sampat, 2009) as it reflects an economic decision of the owner of the patent. Thus, patent renewal can be thought of as revealing the private value that the owner attributes to the patent.

Estimating patent renewal is not a trivial endeavor as patents can be maintained for only a small number of years, with different payments charged by the patent office each year. Addtionally, Pakes (1986) highlighted a real option approach by considering that renewing a patent does not only extend patent protection for a limited time but also provides the option of future extensions. This approach has been extended and modelled by Maurseth (2005) using survival analyses. Patent renewal can also be modelled using binary approaches, by considering for each patent whether it had been renewed up until that benchmark or not.

Each approach has its advantages: binary analyses provide an easy to understand framework, with relatively few assumptions, but do not use all information by only assessing a single timeframe. Survival analyses better model the renewal process, incorporating the full variation in the renewal indicator, but must rely on different and stronger assumptions on the underlying model. The main assumption being that patents with different citation rates have proportional hazard functions. For our analyses, this would restrict patent citations to have effects on the patent renewal decision that are constant over time. This assumption is generally violated in our data as will be shown in the analyses at the end of the paper. For this reason, the survival analyses will depict an average hazard, which may be harder to interpret economically. Finally, a linear model where renewal time is directly estimated can also be used. While neglecting the likely non-linearity, in the relation between patent renewal and patent value, it does produce an easy to understand model with respect to renewal time.

In this paper, we will present survival analyses as they give a more comprehensive picture and allow for deeper inspection of the data. However, in appendix C, we will present alternative methods such as logistic analysis from Hegde and Sampat (2009), and censored linear, i.e.Tobit, analysis as shown in Bakker (2017), to corroborate the main findings from the survival analyses.

Controls

Patents, as the inventions they represent, have a high degree of heterogeneity, and therefore suitable controls need to be added. Thereupon, controls will be added for the main causes of heterogeneity that may affect the citations to a patent and its value. These are: the year of filing, technology, applicant, and value characteristics. In the remainder of this section we will discuss the reasons for the inclusion of each of these controls and how the controls will be implemented.

As is common in patent research, dummies were added for both the year and technological area (IPC3)[7] in which the patent was filed. In some of the robustness tests, the IPC3 level is found too generate to many control variables, which impacted the analyses negatively[8]. In these cases, we resorted to using controls using the 35 FHG categories as defined by Schmoch (2008). These FHG categories are derived by clustering classes from the IPC system and are therefore likely to be good representation of the information in the IPC3 controls. Most patents have multiple IPC3 classes, while some patents cover several FHG categories. In these cases, the dummies have been normalized to reflect a partial count to estimate the average effect the technology classes/categories have on patent renewal, and thereby better controlling for their contribution.

Furthermore, additional controls were added to reflect the attributes of the applicant of the patent[9]. The controls are included because we assume that different applicants likely write different kinds of patent documents, thus affecting their citation rates, and also have different evaluation criteria of the renewal decision of a patent. Moreover, smaller applicants may qualify to pay lower maintenance fees at the USPTO (2016). Additionally, Lampe (2012) found that foreign applicants may have a different citation behavior. Accordingly, as controls we included the following: the type of applicant (i.e. company, government, hospital, individual, university, or unknown/other); the size of the applicant; the experience of the applicant; and the residence country of the applicant. We also included a dummy indicating whether there were multiple applicants. Whenever there were multiple applicants we observe this in a dummy variable and adjust the applicant related variables in the following way: for continuous variables, we defaulted to the oldest and largest applicant as we assume that it is the most experienced actor that will decide on patent renewal; as for the categorical attributes (i.e. applicant type and country of residence), we used a partial count in these cases.

The main difficulty with research involving patent citations and private value is that patent citations, in general, correlate with the quality of the patent (Lanjouw, 2004). Therefore, to distinguish the productive and market value, the

[7] The second highest level of aggregation of the International Patent Classification system, also known as the class level. In our data we observe 128 distinct categories.
[8] For example, in some binary analyses non-convergence was encountered due to the considerable number of dummies which had non-zero values only for a few patents.
[9] For 7 patents no applicant was identified, these applications were not included in the analysis.

general quality of the patent needs to be controlled for. In this paper, we add several controls that have been known to correlate with patent quality and patent value, following the overview by Squicciarini et al. (2013). These are: whether the patent is triadic; the number of countries the patent family members are filed in (an adapted version of Harhoff et al., 2003); the number of claims; and the grant lag of the patent (Harhoff and Wagner, 2009; Régibeau and Rockett, 2010). We validated the inclusion of these control variables in appendix A.

Additionally, we added controls that reflect the broadness and complexity of the patent, since simple and narrow patents may differ from complex and broad patents. We control for complexity using the number of distinct IPC3 classes, and for broadness using the total number of IPC classes present on the patent. Further controls for complexity are the following: the number of backward citations Harhoff et al. (2003); the Trajtenberg et al. (1997) originality; and the Shane (2001) radicalness indicator at the IPC6 level. Here the IPC6 level was chosen as it represents a more fine-grained control than the IPC3 level that is used to control for technology classification. We did not add the number of non-patent references of a patent as an indicator, since it did not provide significant coefficients in the survival regressions and lead to non-convergence in some of the binary analyses. Survival analyses with the indicator included showed that the relevant coefficients for our analyses were not substantially altered because of this omission.

Descriptive statistics

In table 2, the descriptive statistics are listed for the variables that are used in this paper. The descriptive statistics, except for EPO specific indicators, refer to the statistics of USPTO patents. The statistics for most variables differ little as the EPO patents belong to the same DOCDB patent family and therefore have the same technological content.

Because of censoring issues, it may happen that patents are present in one patent office and yet have no family member in the other, because the family member was filed outside the time limits of our sample. It also rarely happens that a DOCDB patent family has multiple granted family members in the same office, which may occur because of divisions of patents. This explains the different number of observations in both EPO and USPTO analyses.

Finally, there is sometimes the problem that no patents belonging to applicants of particularly small nationalities, and/or technology categories, were licensed/sold/maintained. To prevent these patents from falling out, two left-over categories were created for groups containing less than 50 patents: one for applicant nationalities; and one for technology categories. This ensures that we can still correct for patent classes and applicant nationalities, even if there are very small groups created by these controls. However, for the binary analyses regarding licenses and transfers it was necessary to still drop some of the smaller categories, leading to slightly fewer observations in their analyses.

Table 8: Descriptions and descriptive statistics of the continuous variables used in this paper. Variables with # default the largest and oldest applicant if a patent has multiple applicants.

Name	Description	N	Mean	Standard deviation	Min	Max
Dependent variables						
USPTO renewal	The number of years for which maintenance fees have been paid to the USPTO	571816	14.55	5.78	4	20
EPO renewal	The number of years for which maintenance fees have been paid to the USPTO	547365	13.08	4.82	2	20
EPO Opposition	Whether an opposition procedure is instigated against an EPO patent	538261	0.058	0.234	0	1
EPO Licensed	Whether an EPO patent is registered as licensed	538261	0.010	0.100	0	1
USPTO Licensed	Whether an USPTO patent is registered as licensed	569758	0.008	0.087	0	1
EPO Transfer	Whether a sale is registered for the EPO patent	538261	0.326	0.469	0	1
USPTO Transfer	Whether a sale is registered for the USPTO patent	563413	0.340	0.474	0	1
Time controls						
Application year	Dummy for the year in which patent was applied for.	571816	1992.61	5.38	1981	2000
Applicant Controls						
Applicant experience#	Years between the filing date of the current patent and that of the first application filed by the applicant.	571816	36.24	31.15	0	146
Ln(Applt. size) #	Logarithm of the total number of patents ever filed by the applicant, observed over the entire database.	571816	7.39	3.32	0	13.0
Co-patented	Dummy indicating if the patent has more than 1 applicant.	571816	0.06	0.24	0	1
Patent quality controls						
Nr. Countries	Number of distinct patent offices in which the DOCDB family of the patent has at least 1 application present .	571816	7.08	4.12	2	51
Triadic	Dummy to indicate if the DOCDB patent family of the patent also contains a Japanese patent	571816	0.71	0.45	0	1
Backward citations	The number of patents cited by the patent	571816	11.88	12.67	0	198
Originality	The score of the Trajtenberg originality indicator on the IPC6 level	571816	0.51	0.32	0	1
Number of claims	The number of claims registered in the patent	571816	14.57	12.37	0	596
No claims registered	Dummy to indicate patents where the number of claims is unknown. For these patents the number of claims is set to 0	571816	0.0004	0.02	0	1
Grant lag	The number of days between the filing of the application document and the grant date of the patent	571816	798.65	430.99	0	9827
Number of IPC classes	Number of IPC technology classes that are assigned to the patent	571816	5.11	5.08	1	166
Number of distinct IPC3 classes	Number of distinct IPC3 classes that are assigned to the patent	571816	1.74	0.93	1	16
Shane radicalness	The score on the Shane radicalness indicator at the IPC6 level	571816	10.05	11.31	0	198
Citation variables						
EPO self-citations	Number of times a patent has been cited by EPO patents that have the same applicant as the cited patent	547365	0.16	0.36	0	1
EPO other citations	Number of times a patent has been cited by EPO patents that have a different applicant as the cited patent	547365	1.27	2.47	0	126
USPTO self-citations	Number of times a patent has been cited by USPTO applications that have the same applicant as the cited application	571816	1.52	5.65	0	1057
USPTO other citations	Number of times a patent has been cited by USPTO applications that have a different applicant as the cited application	571816	13.26	24.91	0	2792

The main model of this paper uses various categorical variables (IPC categories, applicant country, and sector allocation). Other models replace the IPC classification with FHG categories due to the considerable number of controls created by using the IPC3 classifications. The FHG categories were specifically created to group close IPC classes and seem therefore a decent alternative to control for technological variety. To better understand the categorical variables we display some characteristics of the diversity between them (see table 3).

Table 9: Overview of categorical variables used in this paper. All variables use a partial count. A Herfindahl index is provided for USPTO patents; the values for EPO patents are very similar.

Variable	definition	Number of categories	Herfindahl index
IPC3	Dummy variable to indicate if the IPC3 class (e.g. A01) is present in the patent application.	121	0.035
FHG	Dummy variable to indicate if the patent application is classified in a particular FHG class as determined by Schmoch (2008).	35	0.036
Applicant type	Type of applicant: Company, government, hospital, individual, university or unknown/other.	6	0.92
Applicant country	Country in which the applicant resided at time of filing the patent.	53	0.190

The categories of applicant type and applicant country have a relatively high Herfindahl index, which is due to the presence of a category in which most of observations fall. Most patent applications are, unsurprisingly, filed by companies (96%), while a substantial number of applicants come from the US (32%). These findings entail that the results for patents belonging to non-firm applicants may be substantially different because of their small share in our sample. This is less a concern for patents of non-US applicants as they still represent a majority.

Using the current value of the renewal payments at the EPO[10] and the USPTO, we can determine the variance and average value of the renewal payments in our dataset. Because EPO renewal data is censored, due to the long time it takes to obtain a full picture of the maintenance payments (i.e. 20 years) we also include estimates for a restricted sample for which all maintenance data is available. In the full sample, more patents will be fully maintained as a decision of the owner to abandon the patent at a later stage in the patent life time, will not have been recorded.

[10] The maintenance payments at the EPO are dependent on the national offices at which the patent is registered. Here we will use the pre-grant payment scheme, as defined by the EPO, as an approximation for these payments.

The cumulative values of renewal payments are listed in table 4. This table shows the frequent problem with renewal as an indicator of private value: it doesn't capture extreme values. Renewal payments are useful to distinguish between patents that are valued somewhere between a low bound, after taking filing costs into account, and several tens of thousands of euros/dollars. Extreme outcomes that may derive from radical and novel innovations (e.g. see Verhoeven et al., 2016) will not be captured by this estimation. Fortunately, for our sample, this is not a problem as many patents are not fully renewed. Therefore, we can conclude that the results of this analysis will be mainly relevant to understand the use of the average run-of-the-mill patent, rather than radical or very novel patents.

Table 10: Estimates of the overall private value and variance within our sample.

Name	Description	N	Mean	Standard deviation	Min	Max	% fully renewed
EPO total payment(full sample)	Approximated amount of maintenance/renewal fees payed (€) to maintain the EPO patent	547365	13467	7071	0	24090	34%
EPO total payment (restricted sample)	Approximated amount of maintenance/renewal fees payed (€) to maintain the EPO patent, for patents filed before 1993	256559	13381	7890	0	24090	19%
USPTO total payment	Total amount of maintenance/renewal fees payed ($) to maintain the USPTO patent	571816	7753	5021	0	12600	49%

Validation

In the theory section of the paper, we established 4 hypotheses, which are instrumental to verify whether our assertions that distinguish between productive and market value hold. It is important to note that we assume that owners estimate these values independently of the patent citations. In this view, patent citations reflect value, but do not influence it. In this section, we will test the four hypotheses, using the data and methods outlined in the previous section, to determine whether the proposed framework is reflected in our data.

Productive and market value as different constructs

We determined that productive and market value should measure different constructs. This entails that the correlation between self- and other citations should be relatively small. However, it is still reasonable to assume that the constructs correlate, since technologies that are valuable to their owner are likely to be valuable to others. We estimated the correlation between self- and other citations for both EPO and USPTO patents. The results are shown in table 5.

Table 11: Correlation between self- and non-self-citations for EPO and USPTO patents. *** indicates p<0.001.

Patent source	Nr. observations	Correlation	Squared correlation
EPO	547365	0.32***	0.10
USPTO	571816	0.21***	0.04

It appears that self- and non-self-citations have a reasonably large correlation. This correlation is nonetheless, low enough that one indicator only explains a fraction of the variation, as measured by the squared correlation in the other indicator. We therefore conclude that self- and other citations appear to measure different constructs.

Patent citations and market operations

The productive value and the market value of a patent should influence the behavior of the patent owner regarding market operations, such as licensing and selling. It is expected that patents with a high market value are more likely to be licensed or sold. Furthermore, we expect that patents with a higher productive value, i.e. more other citations, are less likely to be sold.

We test the relation between received citations and the probability of sales and licenses using logistic models and a standardized loglinear form of self- and other citations. This is done for both EPO and USPTO patents regarding sales, and only for EPO patents regarding licenses, due to data issues. In these analyses we correct for technology class using the FHG35 classification scheme, which divides patents in 35 industrial classes based on their IPC

classification (Schmoch, 2008). This is used instead of the IPC3 dummies of later analyses, since including these IPC3 dummies lead to many excluded categories of technologies, as well as poor convergence of the logistic analyses. The analyses are shown in table 6:

Table 12: Logit analyses if a patent is licensed or sold for granted EPO or USPTO patents, explained by the patents self- and other citations. Robust standard errors in parentheses, †p=0.05, * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) License registered (EPO)	(2) Transfer registered (EPO)	(3) License registered (USPTO)	(4) Transfer registered (USPTO)
Ln(1+Self-citations) (standardized)	0.0315† (0.016)	-0.0688*** (0.0035)	0.0836*** (0.014)	-0.171*** (0.0034)
Ln(1+Other citations) (standardized)	0.110*** (0.016)	0.103*** (0.0035)	0.303*** (0.018)	0.215*** (0.0036)
Constant	-2.636* (1.04)	-3.135*** (0.68)	-18.22*** (5.03)	-3.136*** (0.76)
FHG35 Dummies	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Applicant controls	Yes	Yes	Yes	Yes
Quality controls	Yes	Yes	Yes	Yes
N	546176	547365	569758	571816
Pseudo R^2	0.202	0.100	0.255	0.115
Log-likelihood	-24694	-310785	-18882	-324367
Recall (%)	0.29	30.16	2.25	37.93
Specificity (%)	100.00	90.67	99.98	87.41
Precision (%)	50.00	60.97	44.09	60.74

The analysis presented provides support for the hypothesis that non-self-citations correlate positively with the chances of a patent being licensed and the chances that a patent is sold. Our other hypothesis, which states that sales are negatively correlated with self-citations and positively correlated with other citations, is also supported by the analyses. All relevant coefficients are significant, which is a result of the large volume of sold patents in our sample. It is interesting to note that the standardized coefficient of self-citations is, in order of magnitude, close to the one related to other citations. This indicates that owners take a substantial interest both in the market value as well as the position of the patent in their IP strategy.

The coefficient regarding self-citations in the analysis of licensed EPO patents is on the edge of significance (p=0.050). This may be related to the functional form as analyses with a linear relation for self-citations are strongly significant (p<0.001). In all other analyses of this paper the functional form was also tested, but this did not affect significance nor the direction of the effect. The reason it does in this analysis is due to the relatively small number of licensed patents (5,537), which makes the effective sample size of this analysis much smaller than any other, presented in this paper. In general, it is also unclear which functional form should be preferred for self-citations since Bakker (2016) only determined the functional form of all citations combined, out of which self-citations are only a minor subset.

The analysis of this section reveals that the framework presented correctly explains the correlations of self- and other citations with patent licenses and patent sales. We also tested if the owner of the citing patent feels directly threatened by the cited patent, which is a major indication of market value since no one will license or acquire a patent if there was no indication of a possible threat (Sherry and Teece, 2004). This analysis is based on reactions to threats by patent opposition and is detailed in appendix B. This analysis confirms that parties, which are threathened by a patent, tend to cite this patent. Thereby confirming that non-self-citations indicate market value.

Patent citations and renewal

Patent owners are more likely to maintain their patents if their value is high. Therefore, we expect that both productive and market value correlate positively with patent maintenance. Following our framework, this entails that both self- and other citations correlate negatively with patent abandonment. To test this we ran a ox survival regression to determine the relation between self- and other citations and patent renewal. Bakker (2016) indicated that citations have a loglinear relation to patent value, therefore we will use a log-linear transform. Moreover, we standardized the resulting indicators to estimate their relative contributions. The resulting analysis is then performed for both EPO and USPTO patents. It bears reminding that negative coefficients indicate a lesser hazard of non-renewal, and therefore a higher value. The results are listed in table 7.

Table 13: Cox survival regressions for the last renewal registered for granted EPO and USPTO patent applications. Citation indicators refer to office counter parts as listed in table 2. Robust standard errors in parentheses, * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) Renewal EPO	(2) Renewal USPTO
Ln(1+Self-citations) (standardized)	-0.0758*** (0.0017)	-0.182*** (0.0021)
Ln(1+Other citations) (standardized)	-0.116*** (0.0018)	-0.199*** (0.0019)
IPC3 Dummies	Yes	Yes
Year dummies	Yes	Yes
Applicant controls	Yes	Yes
Value controls	Yes	Yes
N	547365	571816
Pseudo R^2	0.007	0.009
AIC	9166501	7491354
Log-likelihood	-4583040	-3745463

The measures of productive and market value are both found to be important since they correlate significantly negative with patent abandonment, and thus positively with patent renewal. Hence, the results provide support for the second hypothesis. Furthermore, the analyses show that the standardized coefficients for self-citations are very similar to those of other citations. This indicates that the estimated productive value has a similar explanatory power as the estimated market value. We tested different methods of estimating renewal (Logit and Tobit), this yielded the same results (see appendix C). We also tested the robustness of the SRA approach for EPO, by comparing it to national offices within its territory. The approach was found to be comparable with that of using a single patent office, as is also documented in appendix C.

How do owners value their patents over time?

Patent owners can use their patent for different purposes: protecting their own innovative efforts or monetize them. It is reasonable to assume that patent owners have a different preference towards either of these two goals, and that this value changes over the lifetime of the patent, for instance in relation to the legal status of the patent (Sherry and Teece, 2004). For example, it is possible that owners first file patents in conjunction with a research and development strategy, where patents are used as a method of protecting their invention. Later, when more is known about possible market operations, the attention may switch more into that direction. Alternatively, given that the option value of renewal is diminishing (Pakes, 1986; Serrano, 2010), the amount of protection may be reduced as patent owners start switching to other measures of IP protection. This may decrease the market value as prospective buyers would need to create production and marketing resources, which eats into the time of the patent validity.

We use our established framework to estimate the weight owners give to productive and market value by observing how their respective indicators explain patent renewal at different points during the patent lifetime. The weight itself is estimated by how well each indicator explains patent renewal. As in the previous section, we use standardized versions of the loglinear transformed citation indicators. As such, we can use the size of the coefficients associated with the citations to gauge the importance attributed by the patent owner. This may be informative to understand when patents are more likely to be offered for sale or otherwise monetized. Likewise, it will help to understand the positioning of patents of different ages in a patent portfolio. In the second subsection, we will look at the evolution of the different valuations for patents over time. This analysis will provide a better look on the motivations for maintaining patents over time, and by doing so the expansion of markets for technology.

To evaluate this, we ran Cox regressions with both values, as estimated by patent citations, interacted with the year after the filing of the patent. It is to be noted that there are only a few decision moments for patents at each office. At the USPTO there are decision moments at the 4, 8 and 12-year mark. For EPO patents the renewal decisions are made yearly, but to attain symmetry we also converted this into three distinct periods. This approach also has the benefit of an increased power in the analysis, which leads to smaller standard errors.

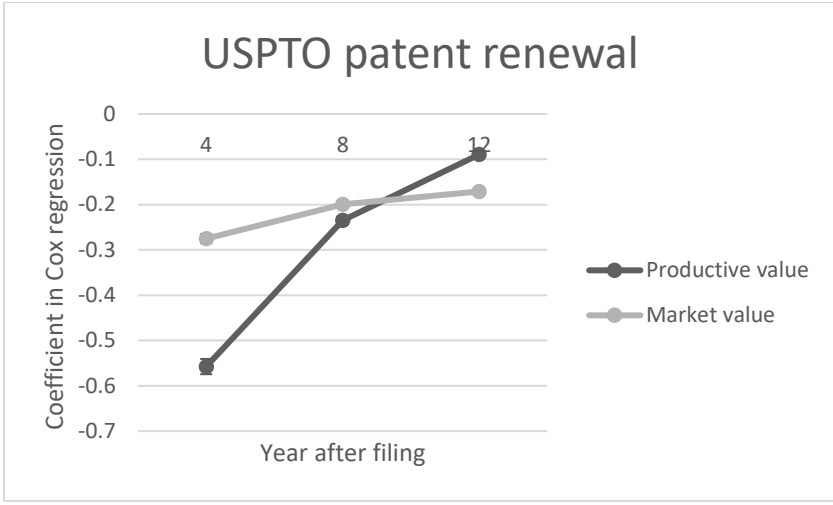

Figure 4: Relative importance of the productive and market value for 3 renewal decisions in the lifetime of USPTO patents. Error bars signal a robust 95% confidence interval.

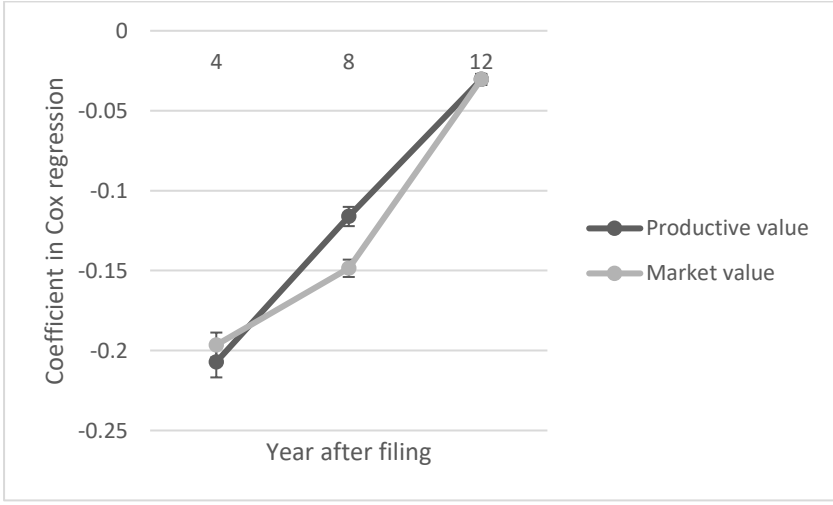

Figure 5: Relative importance of the productive and market value for 3 renewal decisions in the lifetime of EPO patents. Error bars signal a robust 95% confidence interval.

In both figures the drop of the coefficients over time is obvious. This likely is an indication that factors, other than the productive and market value observed in our framework, are becoming more important. These factors may be related to the success of a product or the performance of the firm, and less related to the patent document itself.

The other observation is that while both values become less important near the end of the patent life time, the importance of productive value decreases more than that of market value. This effect appears stronger for USPTO patents than for EPO patents. Therefore, it appears that in the initial stages of a patent's lifetime, owners place a higher importance on the role of the patent in facilitating the production of goods. This focus wanes over time, while owners remain attentive to the interest of other parties in their intellectual property. Because of this process, the complete value of later patents is more likely to be related to their market value and as such they are then more likely to be monetized.

These results also signal measurement issues for those who wish to use patent renewal as a metric to evaluate value. If one only evaluates whether a patent has been renewed in the initial stages (e.g. whether a USPTO patent has been renewed for at least 4 or 8 years) this will capture more productive value than when one considers a full renewal indicator.

However, the EPO analysis may be affected by censoring since we do not observe all renewal decisions for patents filed after 1993 because these decisions were not yet recorded in our dataset. Consequently, we also repeated the analysis with only patents that are filed up until 1993 (see figure 3). This exercise presented a more similar picture to the analysis involving USPTO patents. The divergence between the USPTO and EPO analysis is thus, likely caused by the censoring at the end of the sample.

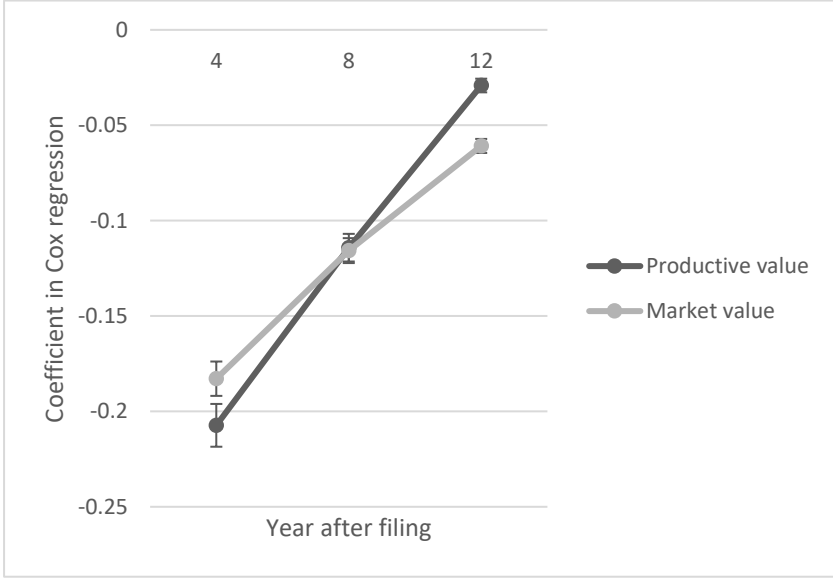

Figure 6: Relative importance of the productive and market value for 3 renewal decisions in the lifetime of EPO patents filed in or before 1993. Error bars signal a robust 95% confidence interval.

Conclusion and discussion

The analyses in this paper reveal that patent owners care about both the productive and market value of their patents when making decisions regarding the patent. Some of the results had already been found in other scientific works. For instance, using a simpler methodology, Thomas (1999) found that self-citations and other citations relate positively to patent renewal. This paper adds to this literature by introducing a new consistent framework which explains directly why certain correlations are expected even when controlling for the technological value of the patent. The paper also adds new insights by consistently finding that self-citations correlate negatively with patent sales, but not with patent licenses. The directionality of the framework is then also validated by using an analysis which ties patent citations to the parties involved in the opposition of patents. Our analyses further reveal that at first, the productive value is important, while for later decisions both productive and market considerations may be weighed in a similar fashion.

Another main contribution is relating patent citations to patent value, by using their function in the patent system. Thus, leading to the conclusion that patent citations indicate value as they indicate potential monetization options for the patent owner. This point of view is a good addition, and perhaps occasional replacement, to the patent citations as knowledge flows narrative, and may be of more use to those interested in the management of intellectual property. For example, patent owners can gauge the market value, as well as potentially interested parties, by observing the received citations. On the other hand, those owners may also want to pay attention to the patents they cite if they want to avoid accidental infringement. The productive and market value are also likely to correspond to defensive and offensive patents respectively. Hence, the citation structure may reveal the positions of parties in complicated technology markets, in a similar vein to the patent thicket analysis by von Gravenitz et al. (2011). Finally, researchers of open innovation can use the framework as an indication for the openness of the innovation strategy of firms: those that generate relatively many outside citations are likely more focused on the market of intellectual property. Alternatively, the renewal analysis presented at the end may help to gauge the importance firms attach to the market value of their patents.

Next to the main conclusions, this paper also provides more technical contributions for patent research, which we will shortly discuss here. First, it shows that patent value indicators do not correlate well, thus necessitating the inclusion of many variables to arrive at a decent value control. By relating the quality controls to patent renewal, it is validated that the quality controls do indeed indicate value in the way that was expected. As for the patent renewal analysis itself, it was demonstrated that our results are independent of the chosen Cox survival regression technique, and also that the SRA method of Van Zeebroeck (2011) for multiple EPO patent offices gives very similar results to the use of a single patent office.

Considering the results found in this paper, it is recommended that researchers include self-citations separately in their analyses, as they appear to represent a quite different process than other patent citations. It is likely that these different processes interact differently with indicators of value and therefore such an indicator could prove to be useful. Consequently, more efforts should be made to better observe patent value, by including additional variables and developing new indicators.

The analyses were made by observing individual patents. This brings with it several issues, out of which the main concern is that larger firms are likely to make decisions based on bundles of patents. Our framework is still applicable regarding the productive and market value of such groups, but will exclude strategic considerations such as the overlap between patents and the diversity of the portfolio. This may also explain the low explanatory value, sometimes found in our analysis, as patent renewals, sales, and licenses may be not decided on the base of the value of an individual patent, but on the base of a patent portfolio. This may lead to low-quality patents being renewed, sold or licensed, simply because they are part of the portfolio of a successful technology protected by multiple patents. Alternatively, patents of a decent quality may be allowed to lapse because they provide overlapping protection with other patents. Thus, the results of the analyses presented in this paper may have inaccuracies with respect to the exact patent value.

In conclusion, this paper shows the importance and potential of considering the legal context of patents when estimating their value, instead of simply observing the technological advances of patent. It seems desirable that researchers of fields that analyze innovation and IP strategies, such as open innovation, continue research in this direction to arrive at a more complete understanding of (the value of) patents and their relation to innovation.

Appendix A: Validation of control variables

Multivariate statistics

In this paper, we use two sets of controls to control for applicant characteristics and the quality of the patent. We will first examine the relation between the set of controls using correlation analysis and Exploratory Factor Analysis (EFA). We only present an analysis based on USPTO patents for this exercise, as the analysis based on EPO patents gives comparable results. First, we will present the correlations between members of the same set of controls (see tables A1 and A2).

Table A1:Correlations between control variables on the applicant level. N=571,816

	Variable Number	1	2	3
Ln(Applt. size)	1	1		
Co-patented	2	0.03	1	
Applicant experience	3	0.67	0.06	1

Table A2: Correlations between the variables that control for patent quality. N=571,816.

Variable name	nr	1	2	3	4	5	6	7	8	9	10
Nr. Countries	1	1									
Originality	2	.23	1								
No claims registered	3	.00	.01	1							
Triadic	4	.27	.32	.00	1						
Number of distinct IPC3 classes	5	.13	.58	.00	.20	1					
Number of IPC classes	6	.33	.54	.00	.27	.44	1				
Backward citations	7	.10	.04	.00	-.01	.06	.10	1			
Number of claims	8	.08	.04	-.02	-.01	.04	.08	.24	1		
Grant lag	9	.07	.08	.00	.01	.07	.11	.18	.14	1	
Shane radicalness	10	.16	.18	.00	.01	.19	.22	.70	.18	.17	1

These tables show that substantial correlations exist between some of the control variables but, with a few exceptions, these do not exceed 0.50. Variables that show high correlation between them are for instance: the variables firm experience and firm size. This is expected because older firms also had more time to file patents and are therefore likely bigger. Another example is the number of distinct IPC3 classes which has a large correlation with the total number of IPC classes. These variables both correlate well with originality. For the remaining variables, we find low correlations, which is consistent with the findings of Squicciarini et al. (2013).

Next, we performed an exploratory factor analysis (EFA) on all controls. This will help determine if the controls indeed indicate two distinct constructs and whether there is any overlap between them. The (Varimax) rotated factor loadings are shown in table A3.

Table A3: Factors extracted using EFA (principal component factors) with the Kaiser criterion and rotated using the Varimax algorithm. Loadings greater than 0.4 are denoted in bold.

Variable (cum. Variance)	Factor 1 (18%)	Factor 2 (33%)	Factor 3 (46%)	Factor 4 (54%)	Factor 5 (62%)	Uniqueness
Nr. Countries	**0.49**	0.12	-0.23	-0.11	-0.16	0.65
Originality	**0.83**	0.05	0.03	0.04	0.06	0.31
No claims registered	0.02	0.03	-0.01	**0.96**	-0.03	0.07
Triadic	**0.56**	-0.11	0.17	-0.07	-0.23	0.58
Number of distinct IPC3 classes	**0.72**	0.08	0.00	0.06	0.13	0.45
Number of IPC classes	**0.77**	0.14	0.02	-0.01	0.02	0.38
Backward citations	0.00	**0.88**	-0.03	0.02	-0.04	0.23
Number of claims	0.02	**0.46**	-0.06	-0.22	-0.08	0.73
Grant lag	0.08	**0.40**	0.00	-0.06	0.12	0.81
Shane radicalness	0.16	**0.85**	0.01	0.04	0.01	0.25
Ln(Applt. size)	0.01	-0.02	**0.92**	0.00	0.02	0.16
Co-patented	0.04	-0.02	0.03	-0.03	**0.94**	0.11
Applicant experience	0.03	0.00	**0.90**	-0.01	0.01	0.20

From these loadings, we can deduce that 3 constructs are measured, as well as two separate control variables. The latter can be observed in factors 4 and 5, which only have high loadings on one variable each: missing claims and co-patented. This is a strong indication, in conjunction with the correlation tables presented before, that the other control variables do not explain whether a patent has missing claims in the database, or if the patent is co-patented.

For the other variables, we detect one factor that appears to be related to the characteristics of the applicant of the patent. One of the two remaining factors appears to be related to the broadness of the patent, while the other appears to be related to its backward citations. Controls that represent the expected value at the moment of the filing of the patent, such as grant lag, claims, and family size, have reasonable loadings on these factors, but retain a high uniqueness. Therefore, we can conclude that the controls that measure applicant behavior and patent quality are indeed measuring different

constructs. As for the quality controls, it is surprising that the two constructs they appear to measure are not directly related to patent value but to the broadness of the patent, as well as the nature of its backwards citations.

Patent indicators and renewal

The models that we will present in this paper will feature the inclusion of many control variables. These controls are included because it is expected that they may influence the relation between patent citations and patent value. In the paper, we have included controls that should represent the general inventive value and quality of the patent. Based on the literature, certain relations between the control variables and patent renewal are expected. By observing these relations in a model without patent citations, we can determine whether the variables approximate the expected constructs.

In this analysis, the IPC technological dummies are not presented since different technological fields may have different lifecycles. This makes it likely that different renewal times may not only represent different value estimations by the owner, but also distinct characteristics of the technology. In fact, this is one of the reasons they are included in the main analyses.

Unlike the analyses in the paper, we will provide hazard ratios for the control variables as they are easier to interpret than the raw coefficients, which are provided by the Cox hazard regressions. It bears reminding that the ratio here refers to the hazard of being abandoned. Therefore, variables that increase this hazard are negatively correlated with patent value. The results are presented in table A4.

Table A4: Hazard ratios for Cox survival regression for granted patent applications, using only control variables to estimate renewal. Hazard ratios presented. Robust standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) USPTO Renewal	(2) EPO Renewal
Application year		
1981	Reference	Reference
1982	0.983	0.996
	(0.017)	(0.014)
1983	0.999	0.993
	(0.017)	(0.013)
1984	1.004	1.011
	(0.016)	(0.013)
1985	1.006	1.020
	(0.016)	(0.013)
1986	1.005	1.013
	(0.016)	(0.013)
1987	0.969*	0.988
	(0.015)	(0.012)
1988	0.948***	0.960***
	(0.015)	(0.012)
1989	0.929***	0.968**
	(0.014)	(0.012)
1990	0.925***	0.953***
	(0.014)	(0.011)
1991	0.896***	0.950***
	(0.014)	(0.011)
1992	0.876***	0.924***
	(0.014)	(0.011)
1993	0.851***	0.902***
	(0.013)	(0.011)
1994	0.824***	0.807***
	(0.012)	(0.0098)
1995	0.792***	0.759***
	(0.012)	(0.0093)
1996	0.807***	0.740***
	(0.012)	(0.0092)
1997	0.777***	0.705***
	(0.011)	(0.0088)

Table A4 continued

	(1) USPTO Renewal	(2) EPO Renewal
Application year (continued)		
1998	0.812***	0.667***
	(0.012)	(0.0085)
1999	0.999	0.626***
	(0.014)	(0.0081)
2000	1.357***	0.594***
	(0.019)	(0.0078)
Applicant type		
Company	0.948***	0.967**
	(0.011)	(0.010)
Government	1.023	1.028
	(0.018)	(0.017)
Hospital	0.958	1.067
	(0.072)	(0.080)
Individual	1.053***	1.079***
	(0.012)	(0.011)
University	0.968	1.076***
	(0.020)	(0.021)
Unknown	1.093*	1.046
	(0.044)	(0.040)
Other applicant controls		
Ln(applt size)	0.972***	0.979***
	(0.0010)	(0.00093)
Co-patented	1.016	1.012
	(0.014)	(0.013)
Applicant experience	1.003***	1.002***
	(0.000092)	(0.000084)
Value controls		
Number of countries	0.983***	0.971***
	(0.00056)	(0.00053)
Originality	1.019*	1.011
	(0.0085)	(0.0074)
No claims registered	8.928***	0.950
	(0.58)	(0.029)
Triadic	1.004	1.027***
	(0.0047)	(0.0043)
Number of distinct IPC3 classes	0.988***	0.995
	(0.0026)	(0.0024)

Table A4 continued

	(1) USPTO Renewal	(2) EPO Renewal
Value controls (continued)		
Number of IPC classes	1.000	0.999*
	(0.00048)	(0.00051)
Backward citations	0.997***	0.987***
	(0.00023)	(0.00062)
Number of claims	0.994***	0.994***
	(0.00018)	(0.00021)
Grant lag(years)	1.162***	0.901***
	(0.0017)	(0.00094)
Shane radicalness	1.005***	1.008***
	(0.00025)	(0.00041)
IPC3 dummies	Significant	Significant
Country dummies	Significant	Significant
N	571816	547365
Pseudo R^2	0.007	0.006
AIC	7508163	9174499
Log-likelihood	-3753870	-4587041

The hazard ratios with respect to the year of patent filing, show that patents of later years are more likely to be renewed. Only in the USPTO analysis for the last years is this trend reversed. This may signal a data issue (i.e. renewal payments from 2012 not yet registered in the database), or a changing of the financial incentives for renewing patents in the 2000s, which may be caused the financial crisis starting in 2007. Interestingly, we do not observe similar shocks for the EPO year dummies. This is likely because EPO renewal is determined much more often, which entails that shocks affect more patents. The observed shock for the last year of USPTO patents is, however, still cause for concern. Therefore, we performed robustness tests in all the analyses in this paper, which showed that excluding the last year did not lead to substantially different results.

The hazard ratios belonging to the dummies concerning the type of applicant, reveal that companies are most likely to maintain their patents, which is unsurprising as they receive the most direct incentives from having patents. On the other hand of the spectrum, individual applicants are less likely to renew their patents. This could be due to the limited resources they can spend on maintenance payments. The other applicant types do not appear to have different renewal characteristics from each other: their hazard ratios are not significantly different. It is also found that larger and newer applicants are more likely to maintain their patents. The latter is surprising as newcomers are maybe more likely to cease their production activities.

The quality controls behave as expected, with commonly used indicators of private patent value, i.e. the number of countries and the number of claims, all being negatively related to the hazard of abandonment. Interestingly, the speed of granting a patent appears to be negatively related to EPO renewal, this may indicate that granting speed as a measure of private value may be better suited for USPTO patents than for EPO patents. The coefficient for triadic patents is not significant at the USPTO, which is probably due to our sample selection in which all patents already have at least a USPTO and EPO family member. Variables that signal complexity show a more diverse picture, with the variables Shane radicalness and originality indicate a higher hazard of abandonment, while the number of distinct IPC3 classes and the number of backward citations indicate a lower hazard. Finally, patents where no claims are registered in the EPO PATSTAT database are very unlikely to be maintained.

This exercise shows that the control variables are generally performing in the way they are expected to. Therefore, we believe that we indeed arrive at a model which can reasonably control for the general patent value, which is likely the result of its underlying technology in the models presented in this paper.

Appendix B: Opposition as a response to threatening patents

We validated our framework by correlating citation indicators based on self- and other citations with the probability a patent is licensed, sold, or renewed. We would also like to determine a more direct measure of market value, which tests whether the owner of the citing patent displays an interest in the citing patent. Unfortunately, licenses and sales of patents are relatively rare events, and moreover, it is hard to consistently extract the name of the new owner from the patent data. For this reason, we turn to another indication of market value which is whether an opposition procedure is instigated against a patent.

The opposition procedure at the EPO is a possible method of invalidating a patent. This procedure allows third parties to challenge the validity of any patent that has been granted within 9 months after its grant date. Such an opposition procedure is then used as a method by which third parties can remove low-quality patents from the patent system without the expense of a lawsuit (Harhoff and Reitzig, 2004). Hence, we expect that opposition filings are mainly initiated by parties for which the granted patent would represent a potential threat (Harhoff et al., 2015). Patents that are potentially threatening to others are patents that have a potential market value for their owner. This is because the owner can use the patent to force the threatened parties into license agreements, use legal action against them, or sell the patent to them. Therefore, we use the existance of the opposition procedure as a proxy for the market value of a patent.

If citations are indeed an indication of market value, then we would expect opposed patents to be cited relatively often by patents that belong to the opposing party. To test this hypothesis, we use opposition data from the April 2014 version of the INPADOC PRS augmented EPO PATSTAT database. For all EPO patents between 1980 and 2005 we extracted information on the listed opposition as well as the patents that cited them in the EPO PATSTAT 2013 fall database. From this information, we extracted patents for which the opposing party could be identified as an applicant, by using the HRM aggregated table, for any patent in our database. This lead to the identification of 11528 applicants, of which the vast majority (92%) are identified as companies.

For each of these opposed patents a reference set was created. This set consists of granted, but not opposed patents, filed in the same year and which shared at least one full IPC code with the opposed patent. Full IPC codes are very detailed, and few patents are filed in groups that are identified by these codes: on average 148 possible control patents were found per opposed patent. From this control set, we randomly selected 5 control patents per opposed patent, while discarding any opposed patents for which at least 5 control patents could not be found.

Our matching did not involve any patent quality or patent value considerations, and therefore we control for these characteristics by inserting the same value controls as used in the analyses presented in the main paper. We also control for the number of times the patent has been cited by others than the opposing part. Following Bakker (2017), we model the contribution of patent citations because log-linear as the linear fit causes convergence issues for the logistic models. Finally, there is the possibility that self-citations also play a role in our analyses, therefore the analyses were replicated with a log-linear transformation of them included. Descriptive statistics of relevant variables are listed in table B1.

Table B1: Descriptive statistics of variables used in the analyses of this sub-section.

Variable	N	Mean	Std. Dev.	Min	Max
Patent is opposed	270636	0.17	0.37	0	1
Opposer citations	270636	0.25	1.94	0	353
P(opposer citing)	270636	0.07	0.26	0	1
Ln(non-opposer citations+1)	270636	1.04	1.09	0	7
Ln(self-citations+1)	270636	0.27	0.60	0	6
DOCDB family size	270636	9.45	8.84	1	392
Trajtenberg originality at IPC6 level	270636	0.62	0.28	0	1
No claims registered	270636	0.002	0.05	0	1
Triadic	270636	0.58	0.49	0	1
Number of distinct IPC3 classes	270636	2.03	1.12	1	16
Number of distinct IPC classes	270636	6.93	6.53	1	166
Number of backward citations	270636	5.45	5.12	0	149
Number of claims	270636	13.82	10.27	0	422
Grant lag(days)	270636	1860.15	794.08	0	8826
Shane radicalness at IPC6 level	270636	6.03	6.54	0	147

For the resulting sample, we established whether a patent was cited by a patent that was applied for by the opposing party of the reference patent and if so, how many times[11]. These occurences are then primarily explained as a function of the dummy that indicates if the patent was opposed, or part of the reference group. The results of this exercise are listed below in table B2.

[11] An applicant may have multiple patents citing the opposed patent.

Table B2: Regressions estimating the chances of being cited by the opposing party of an EPO opposition procedure for opposed patents and their controls. Robust standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) Logit P(opposer citing)	(2) OLS Opposer citations	(3) Poisson Opposer citations	(4) Logit P(opposer citing)	(5) OLS Opposer citations	(6) Poisson Opposer citations
Patent is opposed	1.336*** (0.017)	0.505*** (0.017)	1.228*** (0.031)	1.342*** (0.017)	0.499*** (0.017)	1.218*** (0.032)
Ln(non-opposer citations+1)	0.691*** (0.0067)	0.261*** (0.0084)	0.828*** (0.014)	0.562*** (0.0088)	0.155*** (0.0075)	0.643*** (0.020)
Ln(self-citations+1)				0.358*** (0.013)	0.351*** (0.022)	0.382*** (0.026)
Value controls	Significant	Significant	Significant	Significant	Significant	Significant
Constant	-3.125*** (0.031)	-0.0835*** (0.023)	-2.749*** (0.073)	-3.039*** (0.032)	-0.0482* (0.022)	-2.580*** (0.069)
N	270636	270636	270636	270636	270636	270636
R^2		0.034			0.043	
Pseudo R^2	0.146		0.224	0.153		0.238
AIC	120869	1116325	380535	119883	1114009	373297
Log-likelihood	-60422	-558149.4	-190255	-59928	-556991	-186634

From these results, we observe that opposed patents are more likely to be cited by patents belonging to the party that opposes the patent. This holds even when controlling for the quality, as well as the number of other citations received by the patent. This effect is, besides highly significant, also substantial: the binary analysis estimates that the odds ratio is about 3. Moreover, the linear regression indicates that, on average, the opposed patent receives 0.51 more citations than the patents selected for control. The Poisson regression confirms that this increase is significant. Finally, this was also verified by a negative binomial regression which gave a very similar coefficient (not shown here).

We conclude that opposed patents have an increased probability to be cited by the opposing party. This is consistent with our hypothesis that patent owners feel threathened by patents that they cite. From the previous discussion, we infer from this result, that non-self-citations are indicative of market value.

Appendix C: Evaluating the robustness of our renewal analysis

We approximated the value that owners attribute to their patents, by using Cox survival regressions. However, in that analysis we used several large assumptions, which deserve a closer examination. First, we will evaluate the choice of using survival analyses rather than binary analysis or a censored linear analysis. Second, we evaluate the SRA method of Van Zeebroeck (2011) by comparing the EPO analysis to analyses using only German, French or UK patents.

Using other estimation methods for renewal

In this paper, we employ Cox survival analyses to estimate the relative value of different types of patent citations. As described earlier, this method is used because it takes full advantage of the variance in the data. Yet, other methods may also be considered. For example, binary analysis such as the linear probability model used by Hegde and Sampat (2009). Hence, we ran logistic regressions as well as linear regressions. For the latter, the issue of censoring needs to be considered as patents that are renewed untill their full term may have a larger unobserved value. Here we follow therefore the approach of Bakker (2017) and apply a Tobit regression. The results of this exercise are listed in table C1.

Table C1: Results of other estimation methods to estimate patent renewal. Robust standard errors between parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) Tobit EPO Renewal	(2) Logit Full term EPO	(3) Tobit USPTO Renewal	(4) Logit full term USPTO
Ln(1+Self-citations) Standardized	0.469*** (0.013)	0.119*** (0.0046)	1.444*** (0.017)	0.251*** (0.0033)
Ln(1+Other citations) (Standardized)	0.731*** (0.012)	0.245*** (0.0050)	1.710*** (0.016)	0.335*** (0.0037)
IPC3 Dummies	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Applicant controls	Yes	Yes	Yes	Yes
Value controls	Yes	Yes	Yes	Yes
N	286805	286759	571816	571816
Pseudo R^2	0.019	0.051	0.035	0.107
AIC	1606550	265084	2507529	708236
Log-likelihood	-803070	-132375	-1253551	-354071

The coefficients of the logistic regressions are unfortunately a biased representation of the population coefficient. Allison (1999) found that the logistic coefficients scale with the unexplained variance in the model. Consequently, when there is a different amount of unexplained variance, the exact coefficients cannot be compared between regressions.Fortunately, the bias affects all coefficients by the same factor and therefore, the relation between the coefficients should be similar as the one found in table 5. Therefore, we compare the ratio of coefficients listed in this table to the ones found in this section (see table C2).

Table C2: Ratios between the coefficients of self- and other citations in analyses that estimate patent renewal using different methods.

Office	EPO			USPTO		
Analysis method	Cox survival	Tobit	Logit	Cox survival	Tobit	Logit
Ratio self-citations/ other citations	0.65	0.64	0.49	1.09	1.18	1.33

This comparison shows that the ratio of the estimated coefficients in table C1 remains quite close to that of the survival regressions. Hence, we conclude that the results of the patent renewal analysis are robust with respect to the estimation methods used in this paper.

EPO renewal

As described in the method section, renewal at the EPO is not a straightforward process, considering that owners of patents need to maintain EPO patents at national offices. In this paper, EPO renewal was estimated using the SRA method of Van Zeebroeck (2011). Yet, as this method has rarely been used to estimate patent value, a robustness test is required to explore how this method fares against methods using only a single patent office. Thus, we repeated the analysis presented in table 5 using renewal data of three largest patent offices within the EPO territory. These are the patent offices of Germany (DE), the United Kingdom (UK), and France (FR). Not all EPO patents have been renewed at least once in all three of these offices, and therefore renewal data is not available for the entire sample of patents from the main analysis. This leads to slightly smaller samples in the individual analyses, whose results are listed in table C3.

Table C3: Comparison of Cox survival analyses with single patent offices that subscribe to the EPO, as compared to the EPO analysis of table 2. Robust standard errors between parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(0) Renewal EPO	(1) Renewal DE	(2) Renewal UK	(3) Renewal FR
Ln(1+Self-citations) Standardized	-0.0758*** (0.00171)	-0.0693*** (0.00177)	-0.0632*** (0.00179)	-0.0618*** (0.00183)
Ln(1+Other citations) (Standardized)	-0.116*** (0.00179)	-0.113*** (0.00186)	-0.106*** (0.00188)	-0.110*** (0.00193)
IPC3 Dummies	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Applicant controls	Yes	Yes	Yes	Yes
Value controls	Yes	Yes	Yes	Yes
N	547365	493347	461048	440165
Pseudo R^2	0.007	0.007	0.007	0.007
AIC	9166501	8251160	8012374	7613663
Log-likelihood	-4583040	-4125370	-4005977	-3806622

The difference found is small: the difference between the coefficient of the SRA regression and the smallest coefficient is 23%, for the coefficient related to self-citations, and 9% for the coefficient related to other citations. Thus, we can conclude that the SRA method presents a decent alternative to using renewal data from a single European office when working with data concernaing the renewal of EPO patents.

References

Akers, N. J. (2000). The referencing of prior art documents in European patents and applications. *World Patent Information*, *22*(4), 309-315.

Alcacer, J., and Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, *88*(4), 774-779.

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological methods and research*, *28*(2), 186-208.

Amable, B., Chatelain, J. B., and Ralf, K. (2010). Patents as collateral. *Journal of Economic Dynamics and Control*, *34*(6), 1092-1104.

Azagra-Caro, J. M., Mattsson, P., and Perruchas, F. (2011). Smoothing the lies: The distinctive effects of patent characteristics on examiner and applicant citations. Journal of the American Society for Information Science and Technology, 62(9), 1727-1740.

Bakker, J. Van Looy, B (2015). A critical analysis of patent citations as a measure of patent value, paper presented at the DRUID conference, June 2015

Bakker, J., Verhoeven, D., Zhang, L., and Van Looy, B. (2016). Patent citation indicators: One size fits all?. *Scientometrics*, *106*(1), 187-211.

Bakker, J. (2017). The log-linear relation between patent citations and patent value. *Scientometrics*, *110*(2), 879-892.

Barabási, A. L., and Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Belenzon, S. (2012). Cumulative Innovation and Market Value: Evidence from Patent Citations. *The Economic Journal*, *122*(559), 265-285.

Bessen, J. (2008). The value of US patents by owner and patent characteristics. *Research Policy*, *37*(5), 932-945.

Bessen, J., and Meurer, M. J. (2008). *Patent failure: How judges, bureaucrats, and lawyers put innovators at risk*. Princeton University Press.

Carpenter, M. P., Narin, F., and Woolf, P. (1981). Citation rates to technologically important patents. *World Patent Information*, *3*(4), 160-163.

Cockburn, I., Kortum, S., and Stern, S. (2003). Are all patent Examiners Equal? Examiners, Patent Characteristics and Litigation Outcomes in Cohen, W. and Merrill, S.(eds.) Patents in the Knowledge $ Based Economy.

Cohen, W. M., Nelson, R. R., and Walsh, J. P. (2000). Protecting their intellectual assets: Appropriability conditions and why US manufacturing firms patent (or not) (No. w7552). National Bureau of Economic Research.

Cotropia, C. A., Lemley, M. A., and Sampat, B. (2013). Do applicant patent citations matter?. *Research Policy*, *42*(4), 844-854.

Criscuolo, P., and Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy*, *37*(10), 1892-1908.

Eisenberg, R. S. (2004). Obvious to whom? Evaluating inventions from the perspective of PHOSITA. *Berkeley Technology Law Journal*, 885-906.

Figueroa, N., and Serrano, C. J. (2013). *Patent trading flows of small and large firms* (No. w18982). National Bureau of Economic Research.

Fischer, T., and Leidinger, J. (2014). Testing patent value indicators on directly observed patent value—An empirical analysis of Ocean Tomo patent auctions. *Research Policy*, *43*(3), 519-529.

Fleming, L. (2007). Breakthroughs and the" long tail" of innovation. *MIT Sloan Management Review*, *49*(1), 69.

Galasso, A., Schankerman, M., and Serrano, C. J. (2013). Trading and enforcing patent rights. *The RAND Journal of Economics*, *44*(2), 275-312.

Gambardella, A., Harhoff, D., and Verspagen, B. (2008). The value of European patents. *European Management Review*, *5*(2), 69-84.

Gay, C., and Le Bas, C. (2005). Uses without too many abuses of patent citations or the simple economics of patent citations as a measure of value and flows of knowledge*. *Economics of Innovation and New Technology*, *14*(5), 333-338.

Gittelman, M. (2008). A note on the value of patents as indicators of innovation: Implications for management research. *The Academy of Management Perspectives*, *22*(3), 21-27.

von Graevenitz, G., Wagner, S., and Harhoff, D. (2011). How to measure patent thickets—A novel approach. *Economics Letters*, *111*(1), 6-9.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools* (No. w8498). National Bureau of Economic Research.

Hall, B. H., Jaffe, A., and Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, 16-38.

Harhoff, D., and Reitzig, M. (2004). Determinants of opposition against EPO patent grants—the case of biotechnology and pharmaceuticals. *International journal of industrial organization*, *22*(4), 443-480.

Harhoff, D., von Graevenitz, G., and Wagner, S. (2015). Conflict Resolution, Public Goods, and Patent Thickets. *Management Science*, *62*(3), 704-721.

Heald, Paul J. "A Transaction Costs Theory of Patent Law." *Ohio St. LJ* 66 (2005): 473.

Hegde, D., and Sampat, B. (2009). Examiner citations, applicant citations, and the private value of patents. *Economics Letters*, *105*(3), 287-289

Heller, M. A., and Eisenberg, R. S. (1998). Can patents deter innovation? The anticommons in biomedical research. *Science*, *280*(5364), 698-701.

Hung, S. W., and Wang, A. P. (2010). Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network. *Scientometrics*, *82*(1), 121-134.

Jaffe, A. B., and Trajtenberg, M. (1999). International knowledge flows: evidence from patent citations. *Economics of Innovation and New Technology*, *8*(1-2), 105-136.

Jafie, A. B., Trajtenberg, M., and Fogarty, M. S. (2000). knowledge spillovers and patent citations: evidence from a survey of Inventors. *NBER/Sloan*, 21.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1992). *Geographic localization of knowledge spillovers as evidenced by patent citations* (No. w3993). National Bureau of Economic Research.

Karkinsky, T., and Riedel, N. (2012). Corporate taxation and the choice of patent location within multinational firms. *Journal of International Economics*, *88*(1), 176-185.

Lemley, M. A., and Sampat, B. (2012). Examiner characteristics and patent office outcomes. *Review of Economics and Statistics*, *94*(3), 817-827.

Lampe, R. (2012). Strategic citation. *Review of Economics and Statistics*,*94*(1), 320-333.

Marco, A. C., Myers, A. F., Graham, S. J., D'Agostino, P. A., and Apple, K. (2015). The USPTO patent assignment dataset: Descriptions and analysis.

Maurseth, P. B. (2005). Lovely but dangerous: The impact of patent citations on patent renewal.*Economics of Innovation and New Technology*,*14*(5), 351-374.

Magerman, T., Van Looy, B., and Song, X. (2006). *Data production methods for harmonized patent indicators: Patentee Name Harmonization*. EUROSTAT Working Paper and Studies, Luxembourg.

Mayergoyz, A. (2009). Lessons from Europe on How to Tame US Patent Trolls. *Cornell Int'l LJ*, *42*, 241.

Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, *49*(1), 93-123.

Narin, F., Noma, E., and Perry, R. (1987). Patents as indicators of corporate technological strength. *Research policy*, *16*(2), 143-155.

Noveck, B. S. (2006). Peer to patent: Collective intelligence, open review, and patent reform. *Harv. JL and Tech.*, *20*, 123.

Pakes, A. (1986). *Patents as options: Some estimates of the value of holding European patent stocks* (No. w1340). National Bureau of Economic Research.

Reitzig, M., Henkel, J., and Heath, C. (2007). On sharks, trolls, and their patent prey—Unrealistic damage awards and firms' strategies of "being infringed". *Research Policy*, *36*(1), 134-154.

Régibeau, P. and K. Rockett (2010), "Innovation Cycles and Learning at the Patent Office: Does the Early Patent Get the Delay?", *The Journal of Industrial Economics*, 58 (2): 222-246.

Sampat, B. N. (2010). When do applicants search for prior art?. *Journal of Law and Economics*, *53*(2), 399-416.

Schankerman, M., and Pakes, A. (1985). *Estimates of the Value of Patent Rights in European Countries During thePost-1950 Period* (No. w1650). National Bureau of Economic Research.

Schmoch, U. (2008). Concept of a technology classification for country comparisons. Final report to the world intellectual property organisation (wipo), WIPO.

Shapiro, C. (2001). Navigating the patent thicket: Cross licenses, patent pools, and standard setting. In *Innovation Policy and the Economy, Volume 1* (pp. 119-150). MIT press.

Serrano, Carlos J. "The dynamics of the transfer and renewal of patents." *The RAND Journal of Economics* 41, no. 4 (2010): 686-708.

Tan, D., and Roberts, P. W. (2010). Categorical coherence, classification volatility and examiner-added citations. *Research Policy*, *39*(1), 89-102.

Thomas, P. (1999). The effect of technological impact upon patent renewal decisions. *Technology Analysis and Strategic Management*, *11*(2), 181-197.

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand Journal of Economics*, 172-187.

Trajtenberg, M., Henderson, R., and Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and new technology*, *5*(1), 19-50.

Sampat, B.N., 2005. Determinants of patent quality: an empirical analysis. Mimeo Columbia University.

Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators Scientometrics, 26 (1993), pp. 193–211

Shane, S. (2001). Technological opportunities and new firm creation. *Management science*, *47*(2), 205-220.

Van Zeebroeck, N. (2011). The puzzle of patent value indicators. *Economics of Innovation and New Technology*, *20*(1), 33-62.

Verhoeven, D., Bakker, J., and Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. *Research Policy*, *45*(3), 707-723.

Which patent citation indicator performs best at approximating patent value?

Jurriën Bakker and Bart Van Looy

Abstract

Patent citations are the most used source for arriving at indicators of patent value. However, various sources and methods lead to a large variation in the resulting patent value indicators. In this paper, we evaluate 4 commonly used patent citation indicators, based on EPO patents, USPTO patents, and DOCDB and INPADOC patent families, regarding their ability to explain patent value as measured by patent renewal. Our findings reveal that INPADOC patent family indicators generally perform better when explaining patent value, except for USPTO patents held by US applicants. When further distinguishing between citations received by the focal patent and its family members, it becomes clear that not all citations convey an equal value. For instance, 'Crowding out' phemona become visible: both the presence of (multiple) family members, and the family member's citations, negatively affect the likelihood of renewal (of the focal patent). Combined, these observations suggest the relevance of adopting a nuanced weighting scheme which considers the family characteristics of the focal patent. Indicators using such a weighting scheme are then computed and shown to produce superior results to any of the commonly used indicators.

Keywords: Patent citations, Patent family, Patent value, Patent renewal

JEL Classification O34

Acknowledgements: The authors wish to thank Dirk Czarnitzki, Otto Toivanen and Koen Frenken for their valuable comments and suggestions.

Introduction

Patent based statistics serve as an essential tool to measure innovation. This is because patents are designed to protect innovative efforts, counting and qualifying them creates thus a reasonable approximation of innovative activity. Additionally, patent statistics are relatively easily gathered, categorized and analyzed. This has contributed to the established status that patent based indicators enjoy as innovation measures.

It may be difficult to capture innovative activity using only patent statistics (e.g. Griliches, 1990), by virtue of the heterogeneous nature of patent documents: some are small incremental improvements, while others represent radical or breakthrough innovations (Dahlin and Beherns, 2005). Therefore, indicators have been advanced that provide insights in the value (sometimes also referred to as quality) of patents (see Squicciarini, 2013 for a brief overview). Of these, patent citations are the most used as a result of their extensive validation (e.g. Carpenter et al., 1980; Trajtenberg, 1990; Harhoff et al., 1999; Hall et al., 2005; Gambardella, 2008), and their longtime availability in the most used patent databases.

However, not all patent citations indicators are equal, since researchers may construct them differently: some scholars (e.g. Trajtenberg, 1990; Hall et al., 2005; Czarnitzki et al., 2011) use data that is based on patents from the United States Patent and Trademark Office (USPTO) - primarily because of the popularity of the National Bureau of Economic Research (NBER) data set (Hall et al., 2001). A second group (e.g. Czartnizki et al., 2011; Hottenrott et al., 2016), uses patents from the European Patent Office (EPO), which are sometimes combined with patents applied through the Paris Cooperation Treaty (PCT), that have an EPO designation. This data is often acquired through the “EPO/OECD citation database”, which is described in Webb et al. (2005). The indicators that are derived from these different practices have been shown to be very different from each other (Bakker et al., 2016).

Additionally, some researchers prefer to aggregate citation data on the level of the patent family (e.g. Gambardella et al., 2008; Graham and Harhoff, 2006; Magerman et al., 2011; Neuhäusler and Fritsch, 2012; Bakker, 2017). Generally, patent families are used to identify patents with equal content (often equivalents) in different patent offices, but they can also be used to group patents thought to be related to the same invention (Martínez, 2010). Aggregating data on the patent family level entails counting citations to all members of the same patent family and, when doing research at the level of the individual patent, attributing this score to an individual member of the patent family. Bakker et al. (2016) have also found that there are major differences when computing patent citation indicators using such a family aggregation, as opposed to using data from an individual office. Nakamura et al. (2015) have likewise found substantially different citation networks, when using patent family aggregation.

In face of these practices, it is unclear which patent citation indicator should be used if one wishes to estimate patent value. Bakker et al. (2016) suggest that using patent family aggregation is preferred if one wishes to create a universal indicator that can be (relatively) easily applied, even when one uses data from a single patent office. Be that is it may, comparability between research papers is not usually the main priority of researchers working on a singular project. Accordingly, it would be useful to set a new standard of patent citation indicators, based on the relation between the indicator and patent value. This should harmonize the research of innovation that relies on the value of patents, as well as improving the accuracy of this measure. The establishment of a preferred indicator will also help reduce the workload for each individual research scholar, as they will not be required to gather data and compute multiple citation indicators. They can instead focus their efforts on obtaining citation data from a sole source.

Therefore, we will investigate the performance of different patent citation indicators, by determining their effectiveness at estimating patent value in a large dataset. We do this by ranking the most commonly used indicators, regarding their ability to explain patent value. By using a large dataset, with a vast number of controls, we will be able to determine general trends which may not be observable in much smaller sets when researchers focus on specific subfields.

In the remainder of the paper, we first introduce the relevant dataset, citation indicators, and relevant control variables. Afterwards, we investigate the extent of native biases in each indicator, and we then proceed to explore which patent citation indicators are best at explaining patent value. Finally, we will discuss possibilities for a weighed indicator, which may perform better than the currently available patent citation indicators.

Patent data and variables

We use the same dataset as Bakker (2017), which is based on the EPO PATSTAT database of October 2013, and renewal data from the spring 2014 EPO PATSTAT database. This data set consists of granted USPTO and EPO patents, filed between 1981 and 2000, thus giving each patent enough time to accrue forward citations.

Two different patent family definitions are used to aggregate patent citations: DOCDB and INPADOC. The DOCDB patent family definition groups patents that share the same technical content (Albrecht et al., 2010) and is, generally, used to group equivalent patents which have been filed in different patent offices. The INPADOC patent family is larger, and groups patents that share a part of their priority picture (Martínez, 2010). It can be thought of as representing the underlying invention, when it is protected by multiple overlapping patents. The INPADOC patent family is therefore, usually (Bakker et al., 2016), larger than the DOCDB patent family.

We restricted the dataset to EPO and USPTO patents that have a granted DOCDB patent family member in the other office. This ensures that analyses of EPO and USPTO renewal are comparable. Additionally, because it involves patents with at least two family members, it ensures that a difference is likely present between the family based citation indicators and the counts at the USPTO and EPO offices. Finally, this type of data ensures that only relatively valuable patents are observed, since having family members in other patent offices is a good indication of patent value (Harhoff et al., 2003).

Patent citation indicators

Patent citations are aggregated on the level of the family or application, in a similar vein as Bakker et al. (2016). To keep the analysis tractable, we also used similar definitions, which denote the citation indicator by the entity that it is based on. For example, the ‘EPO count’ variable counts the citations an EPO patent received from other EPO patents. On the other hand, the ‘DOCDB count’ variable counts the number of times the DOCDB family of the focal patent has been cited by other DOCDB families. Performing this exercise for EPO and USPTO patents, as well as the DOCDB and INPADOC patent families, leads to 4 different citation indicators: the EPO count; the USPTO count; the DOCDB count; and the INPADOC count. These indicators are described in table 1.

Table 14: Definitions and descriptive statistics of the indicators used in this dataset. Statistics for the controls are drawn for the USPTO patents, but are very similar for their EPO counterparts due to the sample composition.

Indicator	Definition	N	mean	Std. dev	min	max
EPO count	Number of citations that are received by EPO patents from other EPO patents	547365	1.63	3.37	0	311
USPTO count	Number of citations that are received by USPTO patents from other USPTO patents	571816	14.77	26.65	0	2802
DOCDB count	Number of citations that are received by the patent, or other members from its DOCDB family, from other DOCDB families	571816	21.83	38.82	0	3146
INPADOC count	Number of citations that are received by the patent, or other members from its INPADOC family from other INPADOC families	571816	26.13	82.54	0	4751

Bakker et al. (2016) showed that patent citation indicators can differ substantially. To investigate this in our sample we computed correlations between the different citation indicators. Two tables are presented relating to either EPO patents (table 2), or USPTO patents (table 3).

Table 15: Correlations between patent citation indicators referring to EPO patents. All correlations are significant at the p<0.001 level. N=547,365

	EPO count	DOCDB count	INPADOC count
EPO count	1		
DOCDB count	0.31	1	
INPADOC count	0.14	0.59	1

Table 16: Correlations between patent citation indicators referring to USPTO patents. All correlations are significant at the p<0.001 level. N=571,816

	USPTO count	DOCDB count	INPADOC count
USPTO count	1		
DOCDB count	0.71	1	
INPADOC count	0.39	0.62	1

From these tables, we observe that the correlations between indicators are significantly positive and vary substantially. The correlations between these indicators are expected to be high, as they should measure the same constructs. However, the correlations are not high enough to conclude that all indicators observe the same information, thus confirming that the conclusions by Bakker et al. (2016) also hold on this dataset.

The correlation between the EPO count and the family indicators is much lower than the correlation between the USPTO count and the family indicators. This is likely due to the substantial number of citations in the USPTO system, which in turn also affect the patent family citation indicators. Therefore, we expect more differences to be observed between the EPO indicator and the family indicators in the analyses of this paper. Likewise, patent family based citation indicators may display similar biases as the USPTO indicator due to this high correlation.

Controls

In all analyses, control variables were added to ensure that the always present heterogeneity of patent documents does not create undesired biases. Consequently, we included technology controls (IPC3), year controls, and applicant controls. The latter are included, since different applicants may write different patents, leading to different forward citation patterns. Additionally, different applicants may also have different filing strategies leading to differences between family and non-family indicators. The applicant controls refer to the size of the applicant (as measured by the total number of their patent applications observed in our database), the age of the applicant at the time of filing (as measured by the number of years between the year of the applicant's first application and that of the currently observed patent), whether the patent had more than 1 applicant, and the nationality of the applicant. In the rare occasions (~6%) that a patent has more than 1 applicant we take the value of the largest and oldest applicant, as we assume that it is this applicant that handles any maintenance decisions. The controls used in our analyses are shown in table 4.

Table 17: Descriptive statistics of the control variables used in this paper. * indicates that a partial count is used to generate dummies. Variables with # refer to maximum values if more applicants are present.

Indicator	Definition	N	Mean	Std. dev	Min	Max
IPC3*	Variable to indicate if the IPC3 class (e.g. A01) is present in the patent application	N/A	N/A	N/A	N/A	N/A
Application Year	Year in which the application was applied for at the patent office	571816	1992.61	5.38	1981	2000
Ln(Applt. size) #	Logarithm of the total number of patents ever filed the applicant	571816	7.39	3.32	0	13.0
Applicant type*	Type of applicant: Company, government, hospital, individual, university or unknown	N/A	N/A	N/A	N/A	N/A
Applicant experience#	Years between filing of current patent and that of the first application filed by the applicant	571816	36.24	31.15	0	146
Applicant country*	Country in which the applicant resided at time of filing the patent	N/A	N/A	N/A	N/A	N/A
Co-patented	Dummy to indicate if the patent has more than 1 applicant	571816	0.06	0.24	0	1

Estimating patent value using patent renewal

The main analysis of this paper centers around using patent indicators to estimate the private value of patents. We choose to use patent renewal for this purpose. The renewal indicator is derived from the fact that patents at the USPTO and EPO need to be maintained by periodical payments from their owners. These payments can then be used as a value indicator, denoted by 'patent renewal' as they entail direct decisions of the patent owner to continue to pay a fee to extend their patent protection. This indicator has been used to approximate private patent value by several researchers (e.g. Hegde and Sampat, 2008; Thomas, 1999; and Bakker, 2017). In this paper, we generally follow the approach of Maurseth (2005), and estimate patent renewal using a Cox survival framework. In effect, we will be observing the hazard of non-renewal, since survival regressions always model hazards, not sucesses. In our case this refers to a patent no longer being renewed.

Patent renewal can be observed for both EPO and USPTO patents, but it will be computed differently for each office. As the USPTO is an ordinary national office, renewal can be readily established by reviewing maintenance payments registered at the USPTO. The EPO, on the other hand, is a regional office which administers patents from various national offices. To establish EPO renewal, we used the single renewal SRA method of Van Zeebroeck (2011),

as explained in Bakker (2017). This metric constructs the EPO renewal time as the longest time a patent has been maintained at one of the national offices that subscribe to the EPO. This allows for a comparable renewal metric for both EPO and USPTO renewal. All USPTO patents had enough time to be fully renewed considering this decision is taken 12 years after the filing date at the USPTO. EPO renewal is censored for patents with application years after 1993, for the reason the last potential EPO renewal decision only takes place 20 years after the filing of the patent. As such, in models using only the full renewal time, we only employ EPO data up until 1993. For the resulting renewal indicators, we present the descriptive statistics in table 5.

Table 18: Descriptive statistics of the renewal indicators used.

Indicator	Definition	N	Mean	Std. dev	Min	Max
USPTO renewal	Latest fee payment registered at the USPTO	571,816	14.55	5.78	4	20
EPO renewal	Latest fee payment at any of the offices that subscribe to the EPO	547,365	13.08	4.82	2	20

The econometric model

In this paper, we set out to find the best citation indicator to use when estimating patent value. And so, our first objective is to find the citation indicator that best explains patent value, as to reduce the noise to signal ratio when this indicator is used as a proxy for patent value. If patent value were to be found as a simple monetary equivalent, this research could be operationalized by computing the shared variance (i.e. squared correlation) between the citation indicators and patent value. This is however not the case, since we rely on patent renewal data.

Therefore, we construct a model that explains patent renewal as a function of patent citations, using the previously described Cox survival regressions. This model will be run several times, with a different patent citation indicator each time. Due to the different indicators the model will have each time a different explanatory power, which is indicated by the fit of the model. The best citation indicator is then revealed by its presence in the model that best explains patent renewal. This process is commonly referred to as horse-race regressions.

Because of the nature of the Cox survival regressions, there is unfortunately no fit statistic that is easy to interpret, such as an R^2. Fortunately, there are other fit statistics that show which model has the best fit. Out of these we chose to focus on the Akaike Information Criterion (AIC) since it can be used to rank models when they are computed on the same dataset[1]. The model with the best fit is then indicated as the one having the lowest AIC score.

[1] The AIC is a linear transformation of the log-likelihood, if the different models have the same number of variables. Therefore, this is equivalent to comparing the log-likelihood of the models.

Unfortunately, this method does not provide an easily understandable estimate of the differences between the various computed models. Consequently, we show binary (logistic) regressions where we use the citation indicators to explain the chances a patent is fully renewed. (available in appendix A). By doing so we can compare fit statistics such as precision, recall and specificity. In appendix B we use a method (i.e. a J-test) which does not rely on finding the best model fit. However, this method works best in a linear setting, which is why we use a model based on censored regression instead of a Cox survival model.

Using patent citation indicators to estimate patent value

Which indicator to use?

Researchers that want to use patent citations to measure patent value can choose from several patent based indicators, and will generally be interested in the indicator that best correlates with patent value. As discussed in section 2 we perform Cox survival regressions to explain patent renewal using the different patent citation indicators. Due to the nature of survival regressions, the models will indicate the hazard of non-renewal.

Additionally, we reviewed how well the USPTO count indicator performs for the renewal of their EPO DOCDB family members, and vice versa. This analysis reveals the efficacy of using USPTO patent data to approximate the value of EPO patents, and vice versa. This may be helpful for researchers that want to estimate patent value but do not have the required citation dataset.

The indicators are presented using a log-linear transformation, as it presents a better model of the relation between private value and patent citations (Bakker, 2017). We added controls for technology and year, as well as several variables to control for applicant characteristics: size, age, country of origin, and applicant type.

In this section, we will cater to the situation in which only one patent citation indicator is used to explain patent value., i.e. the most common situation in innovation research. We will do this by using a “horse-race” approach to rank the patent citation indicators, regarding their ability to explain patent value in our dataset. This entails the computation of separate regressions that use patent citations to explain patent renewal, which differ on the citation indicator that is present. These analyses are presented in table 6 for renewal of EPO patents, and in table 7 for renewal of USPTO patents. We will rank the indicators based on the model fits (as measured by the AIC) of the analyses in which they are present. The indicator that is present in the model with the best fit, i.e. the model with the lowest AIC score, is then deemed to have the best performance at explaining patent value.

Table 19: Horse-race Cox survival regressions to determine which citation indicator best explains EPO renewal. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) EPO Renewal time	(2) EPO Renewal time	(3) EPO Renewal time	(4) EPO Renewal time	(5) EPO Renewal time
Ln(1+EPO count)		-0.218***			
		(0.0024)			
Ln(1+USPTO count)			-0.197***		
			(0.0017)		
Ln(1+DOCDB count)				-0.292***	
				(0.0019)	
Ln(1+INPADOC count)					-0.294***
					(0.0019)
Ln(Applt. size)	-0.0232***	-0.0188***	-0.0192***	-0.0162***	-0.0162***
	(0.00094)	(0.00094)	(0.00094)	(0.00094)	(0.00094)
Applicant experience	0.00257***	0.00254***	0.00229***	0.00194***	0.00183***
	(0.000083)	(0.000083)	(0.000083)	(0.000083)	(0.000083)
Co-patented	-0.00114	-0.00996	-0.000503	0.00520	0.00635
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes	Yes
Country dummies	Yes	Yes	Yes	Yes	Yes
N	547365	547365	547365	547365	547365
Pseudo R^2	0.004	0.005	0.005	0.006	0.006
AIC	9192257.0	9183751	9178888	9168307	9167176
Log-likelihood	-4595931	-4591676	-4589245	-4583955	-4583389

Table 20: Horse-race Cox survival regressions to determine which citation indicator best explains USPTO renewal. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) USPTO Renewal time	(2) USPTO Renewal time	(3) USPTO Renewal time	(4) USPTO Renewal time	(5) USPTO Renewal time
Ln(1+EPO count)		-0.195*** (0.0028)			
Ln(1+USPTO count)			-0.247*** (0.0019)		
Ln(1+DOCDB count)				-0.243*** (0.0021)	
Ln(1+INPADOC count)					-0.234*** (0.0021)
Ln(Applt. size)	-0.0268*** (0.0010)	-0.0224*** (0.0011)	-0.0220*** (0.0011)	-0.0204*** (0.0011)	-0.0205*** (0.0011)
Applicant experience	0.00271*** (0.000092)	0.00264*** (0.0000916)	0.00230*** (0.000092)	0.00211*** (0.000092)	0.00205*** (0.000092)
Co-patented	0.0248 (0.013)	0.0173 (0.013)	0.0266* (0.013)	0.0352** (0.013)	0.0368** (0.013)
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes	Yes
Country dummies	Yes	Yes	Yes	Yes	Yes
N	571816	571816	571816	571816	571816
Pseudo R^2	0.005	0.006	0.007	0.007	0.007
AIC	7519602	75144334	7503404	7505840	7506419
Log-likelihood	-3759599	-3757014	-3751499	-3752717	-3753007

In these analyses, we observe that there is not one indicator that performs best at explaining both the renewal of EPO patents, and the renewal of USPTO patents. The lowest AIC score, which indicates the best fit, is found in the analysis containing the INPADOC count indicator for the sample of EPO patents. For the sample of USPTO patents the lowest AIC score is found for the analysis containing the USPTO count indicator. The DOCDB count indicator appears as a compromise, performing second best at explaining renewal for both the sample of EPO patents, and the sample of USPTO patents. It would thus appear that in the case of EPO patents the addition from other citation sources may prove to be more useful than for the case of USPTO patents.

Interestingly, using USPTO citation data to estimate the value of EPO patents achieves a lower AIC than using EPO citation data. Therefore, we would advise to refrain from using EPO citations to estimate patent value, unless no other option is available.

We tested the results from this section using different model specifications, such as a logit, which estimates the probabilities of full renewal, and a Tobit, which estimates renewal time (see appendix A). From the results of the logistic analyses we observe that using a better suited citation indicator adds about 2 percent points in either the specificity or sensitivity of the estimate. This increase is also close to the lowest improvement, which is observed by introducing any citation indicator. Thus, the effect of choosing a better citation indicator can be to double the additional explained variance gained from introducing a patent citation indicator. We also performed a J-test (see appendix B) which allows for a ranking of citation indicators, confirming the rankings shown in this section.

The effects of applicant origin on value estimates

We discovered that a home bias also exists in patent citation indicators (see appendix C). Accordingly, we should evaluate whether this affects the way in which patent citation indicators correlate with patent value. We have found that for EPO patents the INPADOC derived citations were better at explaining patent renewal, while for the USPTO patents USPTO derived citations proved superior. However, this effectiveness may come at the cost of biased estimates for domestic and foreign patents, in which one or the other is consistently over/under valued, even when appropriate control variables are included.

In this subsection, we will explore this question by repeating the analyses of the previous section, by using a sample split by applicant origin. This will allow us to determine if patent citation indicators will maintain the same ranking we found in the previous section if they are applied on sets of patents that have applicants of different origins. By doing so we can determine if the bias we detected in appendix C affects the effectiveness of the patent citation indicators. We assume that the main difference in the way patent citation indicators approximate value will differ between domestic and foreign applicants, following the results from appendix C and the intuition behind the home bias from Criscuolo (2006).

We first determine which patents are foreign and domestic, by applying the rule that patents which have at least one domestic owner are deemed domestic, while all others are foreign. In this analysis, we will again show the results for all indicators, with exception of the EPO count for USPTO patents and the USPTO count for EPO patents, as these counts are never superior (including in this analysis) and are in general unlikely to be used in this way. The analyses for EPO patents are shown in table 8, and the analyses for USPTO patents are shown in table 9.

Table 21: Cox survival analyses explaining patent renewal of EPO patents with different patent citation indicators for both domestic (EPO) and foreign applicants. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) EPO Renewal time Domestic applicants	(2) EPO Renewal time Domestic applicants	(3) EPO Renewal time domestic applicants	(4) EPO Renewal time Foreign applicants	(5) EPO Renewal time Foreign applicants	(6) EPO Renewal time Foreign applicants
Ln(1+EPO count)	-0.259***			-0.175***		
	(0.0037)			(0.0031)		
Ln(1+DOCDB count)		-0.282***			-0.295***	
		(0.0028)			(0.0025)	
Ln(1+INPADOC count)			-0.287***			-0.294***
			(0.0028)			(0.0024)
Ln(Applt. size)	0.000650	0.000518	0.000981	-0.0238***	-0.0413***	-0.0412***
	(0.0014)	(0.0014)	(0.0014)	(0.0010)	(0.0010)	(0.0010)
Applicant experience	0.00134***	0.00122***	0.00118***	0.00245***	0.00327***	0.00311***
	(0.00012)	(0.00012)	(0.00012)	(0.00010)	(0.00010)	(0.00010)
Co-patented	0.0121	0.0319*	0.0326*	-0.0412*	-0.0633***	-0.0631***
	(0.016)	(0.016)	(0.016)	(0.019)	(0.019)	(0.019)
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes	Yes	Yes
Country dummies	Yes	Yes	Yes	Yes	Yes	Yes
N	244891	244891	244891	302474	302474	302474
Pseudo R^2	0.006	0.007	0.007	0.004	0.006	0.007
AIC	3938991	3934148	3933884	4744438	4733749	4732829
Log-likelihood	-1969346	-1966925	-1966793	-2372069	-2366725	-2366264

Table 22: Cox survival analyses explaining patent renewal of USPTO patents with different patent citation indicators for both domestic (US) and foreign applicants. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) USPTO Renewal time Domestic applicants	(2) USPTO Renewal time Domestic applicants	(3) USPTO Renewal time Domestic applicants	(4) USPTO Renewal time Foreign applicants	(5) USPTO Renewal time Foreign applicants	(6) USPTO Renewal time Foreign applicants
Ln(1+USPTO count)	-0.299***			-0.229***		
	(0.0037)			(0.0023)		
Ln(1+DOCDB count)		-0.241***			-0.247***	
		(0.0040)			(0.0024)	
Ln(1+INPADOC count)			-0.209***			-0.248***
			(0.0038)			(0.0024)
Ln(Applt. size)	0.00283	0.00246	0.00113	-0.0514***	-0.0494***	-0.0479***
	(0.00220)	(0.00220)	(0.00220)	(0.000975)	(0.000976)	(0.00098)
Applicant experience	0.00134***	0.00122***	0.00118***	0.00245***	0.00327***	0.00311***
	(0.00012)	(0.00012)	(0.00012)	(0.00010)	(0.00010)	(0.00010)
Co-patented	0.0121	0.0319*	0.0326*	-0.0412*	-0.0633***	-0.0631***
	(0.016)	(0.016)	(0.016)	(0.019)	(0.019)	(0.019)
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes	Yes	Yes
Country dummies	Yes	Yes	Yes	Yes	Yes	Yes
N	182833	182833	182833	388983	388983	388983
Pseudo R^2	0.009	0.007	0.007	0.005	0.005	0.005
AIC	1653884	165665	1657184	5534006	5533890	5533776
Log-likelihood	-826793	-828180	-828443	-2766853	-2766795	-2766738

The model fits, as indicated by the AIC score, show the following: first, for the renewal of EPO patents the ranking of the indicators remains the same; and second, the ranking for explaining the renewal of USPTO patents does change depending on the origin of the applicants. For patents that have foreign applicants it appears to be better to use the INPADOC patent citation aggregation, while USPTO patents from US applicants are best estimated by using the USPTO citation counts. Therefore, the choice of citation indicator does not only depend on the office of the patent but also on the origin of its applicants.

A combined model of patent citation indicators

We want to know if a single patent citation indicator is dominant, in which case adding another citation indicator does not significantly increase the explained variance in the patent renewal indicator. If this is true, then we would be able to simply recommend to always use that indicator when estimating patent value. Furthermore, we wish to find out if adding multiple citation indicators in one regression leads to an increased performance. And so, we construct a composite regression, in which all citation indicators are present. If all citation indicators have significant coefficients in this regression then they all have added information, and it is thus best to use them all to explain patent value. The results of the composite regressions are listed in table 10.

Table 23: Cox survival regressions using all citation indicators to explain renewal for EPO and USPTO patents. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) EPO Renewal time	(2) USPTO Renewal time
Ln(1+EPO count)	-0.0662*** (0.0028)	-0.0939*** (0.0031)
Ln(1+USPTO count)	0.0732*** (0.0034)	-0.201*** (0.0034)
Ln(1+DOCDB count)	-0.132*** (0.0076)	0.0454*** (0.0078)
Ln(1+INPADOC count)	-0.210*** (0.0065)	-0.0822*** (0.0068)
Ln(Applt. size)	-0.0152*** (0.00094)	-0.0197*** (0.0011)
Applicant experience	0.00185*** (0.000083)	0.00221*** (0.000092)
Co-patented	0.00374 (0.012)	0.0247 (0.013)
IPC3 Dummies	Yes	Yes
Year dummies	Yes	Yes
Sector dummies	Yes	Yes
Country dummies	Yes	Yes
N	547365	571816
Pseudo R^2	0.007	0.008
AIC	9165832	7501852
Log-likelihood	-4582714	-3750720

In both the composite regressions for EPO and USPTO patents all citation indicators have significant coefficients, and hence all contribute significantly in explaining patent value. We also tested this using a J-test, which is a more comprehensive method, and which is detailed in appendix B. This test confirms the results from our composite regression.

In this composite analysis, we observe positive coefficients for some indicators, i.e. the USPTO count indicator in the regression concerning EPO patents, and the DOCDB count indicator in the regression concerning USPTO patents. These coefficients are clearly the result of adding multiple citation indicators in one regression, since their coefficients are always negative when they are the only citation indicator present, as can be seen in tables 6 and 7. This behavior is likely caused by multicollinearity as the two indicators have a correlation of 0.71.

We conclude from this analysis that the best research approach, which would explain patent value best in our large sample, is using all patent indicators at the same time. Be that as it may, such an approach is often unsuitable and has therefore rarely, if ever, been used. Consequently, when trying to explain patent value it is preferred if such a value can be expressed on a single dimension instead of four different ones. This also holds if one wants to use patent citations as a proxy of patent value, since it is likely that not all indicators have the same sign, due to the presence of multicollinearity. Finally, many researchers use small datasets in which it is not feasible to estimate 4 coefficients, especially in the presence of multicollinearity (as in the case in our sample).

Deriving a weighting scheme for patent citations

Set-up

The results from the previous section present a puzzling picture, in which all citation indicators explain patent value, and there exists no single indicator that can be selected to present an optimal indication of patent value. In addition, the indicator rankings, with respect to their ability to explain patent value, depend on the value indicator used (i.e. EPO or USPTO renewal): family based indicators are better at explaining EPO renewal; and the non-family based USPTO count performed better at explaining USPTO renewal. Therefore, it appears that both users of USPTO citation counts and users of family counts estimate patent value best, albeit at different datasets. Even more so, because those relying on USPTO data more frequently try to estimate data of USPTO patents anyway, since USPTO data is often found in a dataset which only covers USPTO patents (e.g. the NBER dataset of Hall et al., 2001).

We believe that a further investigation in the relation between received citations and patent value is necessary. To this end, we split up the received citations in mutually exclusive sources of citations. To accomplish this, we first divided the citations into categories, which depended on the cited entity. This exercise leads to three categories of citations: first, citations made to the patent; second, citations made to its DOCDB family members; and third, citations made to its INPADOC family members, but not to its DOCDB family members. It is to be noted here that such a grouping is possible, since the INPADOC patent family is, in our dataset, almost[2] always larger than the DOCDB patent family.

We then determined the office of the citing patents, leading to the following groups: citations from USPTO patents; citations from EPO patents; citations from PCT applications (i.e. applications that went through the PCT route); citations from patents that are applied at a national office subscribing to the EPO convention (denoted as European patents); and citations from patents from other offices (e.g. the patent offices of Australia, Japan (JPO), Korea (KIPO)). This leads to a combination of 3 cited entity categories and 5 groups of citation origins, which when combined leads to 15 exclusive citation indicators.

[2] This does not hold for 88 EPO patents which appear to be part of the set identified by Bakker et al. (2016), for which the DOCDB family is larger than the INPADOC patent family. For these patents we equated the INPADOC patent family to the DOCDB patent family in our dataset, in order to maintain consistency.

To keep the indicators tractable we denoted them by cited entity-citation origin. For instance, Patent-USPTO refers to the number of citations the patent receives from USPTO patents, while DOCDB-USPTO refers to the number of citations its DOCDB patent family members receive from the same origin. Table 11 lists definitions for each of the citation indicators that are created, by following this procedure to decompose patent citations.

Table 24: Definitions of the exclusive patent indicators that are used in this paper.

Name	Definition
Patent-USPTO	Count of citations made to the patent by USPTO patents
Patent-EPO	Count of citations made to the patent by EPO patents
Patent-PCT	Count of citations made to the patent by PCT applications
Patent-Europe	Count of citations made to the patent by patents from offices subscribing to the EPO
Patent-Other	Count of citations made to the patent by patents from other offices
DOCDB-USPTO	Count of citations made to DOCDB family members of the patent by USPTO patents
DOCDB-EPO	Count of citations made to DOCDB family members of the patent by EPO patents
DOCDB –PCT	Count of citations made to DOCDB family members of by PCT applications
DOCDB –Europe	Count of citations made to DOCDB family members of the patent by patents from offices subscribing to the EPO
DOCDB –Other	Count of citations made to DOCDB family members of the patent by patents from other offices
INPADOC-USPTO	Count of citations made to INPADOC family members of the patent by USPTO patents
INPADOC –EPO	Count of citations made to INPADOC family members of the patent by EPO patents
INPADOC –PCT	Count of citations made to INPADOC family members of the patent by PCT applications
INPADOC –Europe	Count of citations made to INPADOC family members of the patent by patents from offices subscribing to the EPO
INPADOC –Other	Count of citations made to INPADOC family members of the patent by patents from other offices

This approach has two advantages. First, it avoids the double counting that could arise if we would simply use the indicators from the previous section in the same regression. For instance, an analysis containing the USPTO count and DOCDB count indicator would double count any USPTO citations, since the DOCDB count indicator also counts USPTO citations. In fact, the presence of a considerable number of USPTO citations in the DOCDB count is likely the cause of the large correlation between these two indicators. These large correlations may create multi-collinearity problems in our regressions, as can be seen by the coefficients with different signs in table 10. Creating exclusive citation indicators reduces this problem, by avoiding the large correlations between indicators, caused simply by the fact that they partially consist of the same underlying patent data.

The second advantage is that, when the exclusive indicators are added in the same estimation, the coefficients associated with them will represent relative weights. These weights are relative because their exact value depends on the relation between patent citations and patent renewal. The relative weights are then an indication of the relative contribution of each set of citations. If these contributions are similar, they can easily be added together (i.e. using a unitary weight for each contribution) to create a better indicator. If they are dissimilar a weighting scheme may be necessary.

These weights can also be used to test two hypotheses simultaniously. The first hypothesis revolves around the assumption underlying the commonly used patent family based citation indicators, and states that all patent citations are equal and therefore can be simply summed to arrive at an indicator. If this is the case, the weights of the different citations should be equal. The second hypothesis is almost its antithesis, it states that only patent citations made directly to the patent application from patents of its own office contribute to explaining its value. This is implicitly the assumption behind using patents and patent citations only from a single patent office. In this case the only weights associated with the patent citations originating from the same patent office should be non-zero.

Descriptive statistics

The procedure described in the previous section creates several patent citation indicators that are uncommon in the literature. It therefore may be interesting to observe the descriptive statistics related to these indicators.

Table 12 provides descriptive statistics for each indicator used in our analysis. Here we see again the dominance of USPTO citations, as there are many more USPTO citations to USPTO patents than in any other citation category. Additionally, USPTO citations to EPO patents are almost as high as EPO citations themselves. The ‘Other’ category presented in this table is composed of citations that originate from several patent offices, and is discussed in more detail in appendix D.

Table 25: Descriptive statistics for the exclusive patent indicators used in this section. Patent exclusive indicators are split up to describe citations to EPO and USPTO patents. Family indicators can differ substantially because of our exclusive treatment, and are therefore presented for both EPO and USPTO patents.

Indicator name	N	Mean	Std. dev	Min	Max
EPO-EPO	547,365	1.63	3.37	0	311
EPO-USPTO	547,365	1.44	4.21	0	417
EPO-PCT	547,365	0.80	2.20	0	301
EPO-Europe	547,365	0.58	1.36	0	92
EPO-Other	547,365	0.08	0.39	0	20
USPTO-EPO	571,816	1.29	4.47	0	1208
USPTO-USPTO	571,816	14.77	26.65	0	2802
USPTO-PCT	571,816	1.20	4.03	0	1047
USPTO-Europe	571,816	0.31	0.89	0	123
USPTO-Other	571,816	0.18	0.58	0	35
Patent family indicators based on USPTO patents					
DOCDB-EPO	571,816	3.39	8.47	0	1353
DOCDB-USPTO	571,816	6.96	33.14	0	2448
DOCDB –PCT	571,816	2.31	7.23	0	1218
DOCDB -Europe	571,816	1.22	2.40	0	150
DOCDB –Other	571,816	0.91	2.01	0	246
INPADOC-EPO	571,816	0.78	11.06	0	1381
INPADOC -USPTO	571,816	3.42	45.21	0	3037
INPADOC -PCT	571,816	0.65	9.28	0	1163
INPADOC -Europe	571,816	0.13	1.45	0	102
INPADOC -Other	571,816	0.17	2.01	0	95
Patent family indicators based on EPO patents					
DOCDB-EPO	547,365	2.69	8.61	0	1255
DOCDB-USPTO	547,365	17.41	36.78	0	2866
DOCDB –PCT	547,365	2.35	7.43	0	1071
DOCDB -Europe	547,365	0.95	2.27	0	180
DOCDB –Other	547,365	0.97	2.10	0	246
INPADOC-EPO	547,365	1.47	17.59	0	1913
INPADOC-USPTO	547,365	6.57	80.42	0	6163
INPADOC-PCT	547,365	1.34	16.52	0	1610
INPADOC-Europe	547,365	0.29	2.67	0	276
INPADOC-Other	547,365	0.52	3.95	0	262

In appendix D we describe relevant correlations and multivariate statistics. The conclusions of these exercises are twofold: first, there is relatively little correlation between citations received from the various citation sources; and second, there is a very high correlation (in the order of 0.9) between citations received from EPO and PCT documents, which is due to EPO documents often having the same citations as their PCT equivalent. Therefore, we exclude PCT obtained citations from our analyses as to prevent multi-collinearity problems.

The relation between patent citations and patent value

To estimate the relative weights for different citations we employed Cox survival regressions as described in the prevous section. The weights are estimated in a linear rather than a log-linear framework, considering it is possible that the coefficients for certain citation categories are not positive, potentially due to multi-collinearity. This would create problems with a log–linear specification, since taking the log of a negative value returns an imaginary number.

Table 26: Survival regressions to determine the relative weights of exclusive citation indicators when explaining the renewal of USPTO or EPO patents. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001.

	(1) EPO Renewal time	(2) EPO Renewal time	(3) EPO Renewal time	(4) USPTO Renewal time	(5) USPTO Renewal time	(6) USPTO Renewal time
Patent-EPO	-0.0452***	-0.0361***	-0.0361***	-0.0414***	-0.0369***	-0.0369***
	(0.0011)	(0.0011)	(0.0011)	(0.0016)	(0.0016)	(0.0016)
Patent-USPTO	-0.0074***	-0.00396***	-0.00390***	-0.0115***	-0.0105***	-0.0105***
	(0.00080)	(0.00074)	(0.00074)	(0.00019)	(0.00020)	(0.00020)
Patent-Europe	-0.0370***	-0.0264***	-0.0264***	-0.0265***	-0.0175***	-0.0174***
	(0.0015)	(0.0015)	(0.0015)	(0.0027)	(0.0027)	(0.0027)
Patent-Other	-0.0251***	-0.0256***	-0.0255***	-0.0757***	-0.0678***	-0.0678***
	(0.0053)	(0.0053)	(0.0053)	(0.0043)	(0.0043)	(0.0043)
DOCDB-EPO		-0.0213***	-0.0181***		-0.00439***	-0.00389***
		(0.00088)	(0.0013)		(0.00059)	(0.00067)
DOCDB-USPTO		-0.00262***	-0.00346***		0.00177***	0.00161***
		(0.00012)	(0.00020)		(0.000086)	(0.00013)
DOCDB –Europe		-0.0253***	-0.00497		-0.0201***	-0.0102***
		(0.0012)	(0.0035)		(0.0010)	(0.0027)
DOCDB –Other		-0.0479***	-0.0375***		-0.0285***	-0.0307***
		(0.0014)	(0.0035)		(0.0015)	(0.0022)
INPADOC -EPO			-0.00280**			-0.000411
			(0.00094)			(0.00029)
INPADOC – USPTO			0.000867***			0.000170
			(0.00016)			(0.000088)
INPADOC – Europe			-0.0203***			-0.00977***
			(0.0033)			(0.0025)
INPADOC -Other			-0.0102**			0.00216
			(0.0031)			(0.0016)
DOCDB family size	-0.0307***	-0.0191***	-0.0193***	-0.00821***	-0.00615***	-0.00620***
	(0.00048)	(0.00048)	(0.00049)	(0.00039)	(0.00038)	(0.00038)
INPADOC family size	-.00051***	0.0000145	0.000255***	0.000343***	0.000294***	0.000277***
	(0.00015)	(0.000068)	(0.000054)	(0.000032)	(0.000028)	(0.000038)
Ln(Applt. size)	-0.0173***	-0.0164***	-0.0164***	-0.0244***	-0.0233***	-0.0233***
	(0.00091)	(0.00092)	(0.00092)	(0.00090)	(0.00090)	(0.00090)
Applicant experience	0.00226***	0.00197***	0.00195***	0.00237***	0.00233***	0.00233***
	(0.000081)	(0.000081)	(0.000081)	(0.000078)	(0.000078)	(0.000078)
Co-patented	-0.00820	0.00273	0.00248	0.0262*	0.0288*	0.0290*
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes	Yes	Yes
Country dummies	Yes	Yes	Yes	Yes	Yes	Yes
N	547365	547365	547365	571816	571816	571816
Pseudo R^2	0.006	0.007	0.007	0.007	0.007	0.008
AIC	9174624	9164094	9163922	7504012	7502611	7502600
Log-likelihood	-4587108	-4581839	-4581749	-3751798	-3751094	-3751084

The analyses presented in table 13 show that almost all citation sources are significant in explaining patent value. Accordingly, we conclude that only using patent and citation data from one source omits information about patent value. Nontheless, the coefficients that are attributed to different citation sources are very different, with some coefficients being more than ten times larger than others. This is an indication that simply adding all citations together will also not yield an optimal citation indicator.

Interestingly, some coefficients are even positive (and therefore negatively related to patent value), an observation which we determined to be caused, in part by the competition that occurs when there are multiple granted patent family members present in the USPTO system, (see appendix E). This competition is caused by the presence of multiple granted USPTO patents in the same family, which may force the patent owner to choose to maintain some patents, while abandoning others. The added citations to these family members may thus be an indication of an increased competition, as opposed to an increasingly valuable invention. Thus, the existence more citations to the DOCDB patent family members may have a negative relation with the chance that the patent is renewed.

In general, we find that indicators related to citations coming from the EPO system have higher coefficients than those coming from the USPTO system. This reduced impact may give credence to the idea that citations within the USPTO system are less relevant than citations within the EPO system (Michels and Bettels, 2001).

We also find lower coefficients when considering citations to the patent family members, as opposed to the patent itself. This effect is increased if we look at the larger INPADOC patent family, for which some of the coefficients are insignificant, or sometimes even negatively related to patent value. These relations appear to hold to some degree in both patent offices. Therefore, it may be useful to employ a weighting scheme for citations that are received by the patent family.

Creating a better patent citation indicator

The previous section provides weights that can be applied to the various sources of citations when estimating patent value. In this section we will explore whether applying these weights leads to an improved performance when estimating patent value. For this endeavor we create weights out of the estimates, from the preceding subsection, using a procedure in which all weights are determined by dividing the coefficient belonging to the citations by the coefficients belonging to the patent-EPO variable. All indicators that have a negative weight after this procedure, are excluded from the composite indicator. We list the resulting weights of this procedure in table 14.

Table 27: Coefficients from the analysis and the weights derived from them.

Citation source	EPO		USPTO	
	Coefficient analysis	Weight	Coefficient analysis	Weight
Patent-EPO	-0.0361	1.000	-0.0369	1.000
Patent-USPTO	-0.0039	0.108	-0.0105	0.285
Patent-Europe	-0.0264	0.731	-0.0174	0.472
Patent-Other	-0.0255	0.706	-0.0678	1.837
DOCDB-EPO	-0.0181	0.501	-0.00389	0.105
DOCDB-USPTO	-0.00346	0.096	0.00161	0
DOCDB –Europe	-0.00497	0.138	-0.0102	0.276
DOCDB –Other	-0.0375	1.039	-0.0307	0.832
INPADOC-EPO	-0.0028	0.078	-0.000411	0.011
INPADOC –USPTO	0.000867	0	0.00017	0
INPADOC –Europe	-0.0203	0.562	-0.00977	0.265
INPADOC –Other	-0.0102	0.283	0.00216	0

We use the resulting composite indicators to explain patent renewal. This is done using the framework we established for the other patent indicators in the previous section (i.e. we insert the log-linear transformation of the indicator, accompanied by various control variables, in a Cox survival regression which explains patent renewal). This is done for the composite indicators resulting from the USPTO and EPO analyses in analyses that explain USPTO and EPO renewal respectively. The results of this procedure are shown in table 15.

Table 28: Cox survival regressions in which composite indicators explain renewal of EPO and USPTO patents. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001.

	(1) EPO Renewal time	(2) USPTO Renewal time
Ln(1+EPO composite)	-0.332*** (0.0019)	
Ln(1+USPTO composite)		-0.319*** (0.0024)
Ln(Applt. size)	-0.0161*** (0.00094)	-0.0217*** (0.0011)
Applicant experience	0.00194*** (0.000083)	0.00220*** (0.000092)
Co-patented	0.00331 (0.012)	0.0325* (0.013)
IPC3 Dummies	Yes	Yes
Year dummies	Yes	Yes
Sector dummies	Yes	Yes
Country dummies	Yes	Yes
N	547365	571816
Pseudo R^2	0.007	0.008
AIC	9162608	7501239
Log-likelihood	-4581105	-3750417

The AIC scores of the regressions involving the composite indicators, are lower than the lowest AIC score of any regressions from the horse-race analysis of tables 6 and 7. Therefore, we conclude that our weighted indicators are superior to any of the standard indicators that were discussed in this paper.

Conclusion

This paper set out to give more guidance about which patent citation indicators to use, when one wants to approximate patent value. For this endeavor we created four patent citation indicators based on commonly used practices. We then proceeded to test them on their ability to explain patent value, as indicated by patent renewal.

We found that all patent citation indicators explain a significant amount of patent value, as measured by patent renewal. However, due to multicollinearity issues, most researchers will prefer to use a single patent family citation indicator in their research. To aid this choice we performed horse-race regressions to select the patent citation indicator which best explains patent value. This analysis presented a different ranking for patent renewal at different offices. The INPADOC count citation indicator explained most of the variance in the renewal of EPO patents, while the non-family based USPTO count indicator explained most of the variance in the renewal of USPTO patents. Further investigation revealed that the variance in the renewal of USPTO patents, which belong to non-US applicants, is also best explained by the INPADOC count indicator.

Consequently, it appears that whenever the subject of the analysis concerns USPTO patents from US applicants it would be more prudent to use citations to the USPTO patent, rather than a patent family based citation indicator. For EPO patents, and USPTO patents with non-US applicants, patent citations aggregated at the INPADOC patent family level are preferred. These results can be added to the analysis of Bakker et al. (2016), who argue in favor of the INPADOC patent citation indicator, due to the fact that it yields comparable indicators in different citation systems. Therefore, we recommend that the INPADOC family based patent citation indicator should be the standard indicator, when using patent citations to approximate patent value.

To understand why aggregating at the family level does not always bring forth improvements an analysis was performed, which revealed that citations to family members need to be discounted in aggregated counts. This discount may also need to differ when considering patents from different systems. For instance, citations from outside the patent system are relatively more valuable for USPTO patents, while they are relatively less valuable for EPO patents. This may be related to the different citation practices in different system, given that a large amount of outside references to EPO patents are derived from the USPTO and vice versa (Bakker et al., 2016). For instance, the USPTO with its ‘duty of candor’ has many more citations per patent than the EPO, and as such each individual citation may be less relevant to the patent.

This analysis also revealed that patents may suffer from competition if they have other granted family members in the same patent office. In these cases, additional citations to family members may even indicate a lesser value. This is another reason that patent family citation measures cannot be simply implemented, rather they would benefit from the use of a weighting scheme. Consequently, we created two composite indicators, one for each office studied. These composite indicators were then found to explain patent value better than any other of the evaluated patent citation indicators. We therefore recommend that other researchers adopt the weighting schem that lead to these indicators, when they desire a proxy for the value of EPO or USPTO patents.

The results presented in this paper indicate that citations to the INPADOC patent family of a patent are useful indicators of its value. We can see these citations as referring to the underlying invention of the patent, since the INPADOC patent family is designed to describe inventions that are protected by multiple patents. Accordingly, it is likely that the renewal decisions are at least based, in part, on the value of the underlying product.

Nevertheless, we need to place our results within the larger body of literature regarding patent citations in that they explain a significant but small part of the variation in patent value. This can be determined from a closer observation of appendix A, in which patent citations (and a vast array of other indicators) are used in logistic regressions to explain the probability that a patent is renewed to its maximum term. These regressions show that in general only a small part of the variation of patent value is explained, with an even smaller part of this explained by patent citations. Even the best patent citation indicator added only a few percent increase in goodness of fit indicators, such as recall, specificity, and sensitivity. These results agree with the observations of Gay and Le Bas (2005) and should provide a warning for researchers that wish to use patent citations as an approximation of patent value.

In general, this paper shows that choosing a patent citation indicator to approximate innovation is far from an easy decision. But based on the results, this paper advocates for the use of INPADOC aggregated patent citations. Furthermore, when using this indicator, it may be necessary to apply a weighting scheme for which this paper provides an outline. Given the low use of the INPADOC patent family in general, we hope that our conclusions spur more research into this promising construct of innovation.

Appendix A: Other methods to estimate patent renewal

Survival analyses are just one method that can used to estimate patent renewal. Binary methods are also a possibility (e.g. Hegde and Sampat, 2009; Bakker, 2017). These methods have the advantage that the goodness of fit can be assessed in a more intuitive way, by observing the classification of patents regarding their renewal status. Therefore, we produce additional fit characteristics which also allow us to provide an indication of the goodness of fit of our models in general.

To perform this analysis, we computed a logistic estimation regarding whether patents are fully renewed. Unfortunately, this information is not available for all patents, due to censoring at the end of our dataset. This is particularly relevant for EPO patents as we only have full renewal information for patents up until 1993, since the dataset ends in early 2014.

We observe some convergence issues when including patents filed in 1993. This may be due to their renewal information, which may not always have been registered properly, therefore we ran the analysis using only patents up until 1992. The results and ranking of the models remain unaffected considering this data choice, as was confirmed by using simpler controls which allowed the models to converge. We again performed horse-race regressions for all 4 indicators, which are shown in table 16 for the renewal of EPO patents, and in table 17 for the renewal of USPTO patents.

Table 29: Horse-race regressions using a logistic regression method to explain if EPO patents with filing dates up until 1992 will be renewed until their maximum allowed time.
Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) EPO Full renewal	(2) EPO Full renewal	(3) EPO Full renewal	(4) EPO Full renewal	(5) EPO Full renewal
Ln(1+EPO count)		0.459***			
		(0.0067)			
Ln(1+USPTO count)			0.416***		
			(0.0056)		
Ln(1+DOCDB count)				0.617***	
				(0.0065)	
Ln(1+INPADOC count)					0.612***
					(0.0063)
Ln(Applt. size)	0.0206***	0.00886**	0.0142***	0.00878**	0.00813**
	(0.0030)	(0.0030)	(0.0030)	(0.0030)	(0.0030)
Applicant experience	-0.00370***	-0.00348***	-0.00300***	-0.00225***	-0.00189***
	(0.00028)	(0.00028)	(0.00028)	(0.00029)	(0.00029)
Co-patented	0.0206	0.0167	0.0311	0.0164	0.0207
	(0.038)	(0.039)	(0.039)	(0.039)	(0.039)
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes	Yes
Country dummies	Yes	Yes	Yes	Yes	Yes
N	256515	256515	256515	256515	256515
Pseudo R^2	0.018	0.037	0.042	0.058	0.059
AIC	242286	237647	236407	232422	232115
Log-likelihood	-120964	-118643	-118023	-116027	-115874
Recall (%)	0.0231	0.757	1.044	2.797	3.239
Specificity (%)	99.99	99.88	99.81	99.60	99.47
Precision (%)	45.83	58.70	55.27	61.47	58.04

Table 30: Horse-race regressions using a logistic regression method to explain if USPTO patents will be renewed until their maximum allowed time. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) USPTO Full renewal	(2) USPTO Full renewal	(3) USPTO Full renewal	(4) USPTO Full renewal	(5) USPTO Full renewal
Ln(1+EPO count)		0.293*** (0.0040)			
Ln(1+USPTO count)			0.405*** (0.0030)		
Ln(1+DOCDB count)				0.405*** (0.0032)	
Ln(1+INPADOC count)					0.379*** (0.0031)
Ln(Applt. size)	0.0370*** (0.0016)	0.0300*** (0.0016)	0.0299*** (0.0016)	0.0275*** (0.0016)	0.0279*** (0.0016)
Applicant experience	-0.00420*** (0.00014)	-0.00412*** (0.00014)	-0.00353*** (0.00014)	-0.00316*** (0.00014)	-0.00306*** (0.00014)
Co-patented	-0.0403 (0.021)	-0.0293 (0.021)	-0.0365 (0.022)	-0.0530* (0.022)	-0.0555* (0.022)
IPC3 Dummies	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes	Yes
Country dummies	Yes	Yes	Yes	Yes	Yes
N	571816	571816	571816	571816	571816
Pseudo R^2	0.064	0.071	0.088	0.085	0.083
AIC	742404	736838	723126	725915	727005
Log-likelihood	-371039	-368239	-361391	-362779	-363325
Recall (%)	64.20	64.40	64.67	64.72	64.62
Specificity (%)	61.51	62.22	64.03	63.75	63.76
Precision (%)	62.01	62.52	63.76	63.60	63.57

In tables 16 and 17, we observe the same pattern as in the survival analyses: family indicators perform best to explain EPO renewal, while USPTO renewal is best explained by the local office indicator. The logistic regressions further show that EPO renewal is harder to explain than USPTO renewal, which is indicated by the much lower fit characteristics, i.e. the lower recall and precision. A further investigation of the fit characteristics indicated that the EPO analyses estimated that almost no patents were fully renewed. The USPTO analyses produce a more favorable picture with a much higher recall, at the cost of relatively few sensitivity, thus indicating a better fit.

It is not fully clear why the renewal of USPTO patents is better estimated than that of their EPO counterparts. A possibility is that USPTO renewal is decided much earlier and that patent citations may be better at explaining the earlier renewal process. The previous chapter gives some indication regarding the existence of this process. To test this explanation, we denoted EPO renewal only up until 12 years and repeated the analysis (see table 18). Here it is to be noted that, because of the now smaller time window, only EPO patents with application filing dates up until 2000 can be used. Convergence issues unfortunately emerged, and forced us to use 35 FHG controls based on the

classification of Schmoch (2008), instead of the earlier used IPC3 partial count. These FHG groups are derived from the IPC classification and should therefore present a decent alternative to the earlier used IPC3 partial counts.

Table 31: Logit regression on reaching 12 years or more for granted EPO patents until 2000. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1) EPO 12 years	(2) EPO 12 years	(3) EPO 12 years	(4) EPO 12 years	(5) EPO 12 years
Ln(1+EPO count)		0.331***			
		(0.0043)			
Ln(1+USPTO count)			0.290***		
			(0.0030)		
Ln(1+DOCDB count)				0.445***	
				(0.0034)	
Ln(1+INPADOC count)					0.449***
					(0.0034)
Ln(Applt. size)	0.0576***	0.0512***	0.0525***	0.0485***	0.0483***
	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)
Co-patented	0.0577**	0.0704**	0.0622**	0.0503*	0.0490*
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Applicant experience	-0.00537***	-0.00533***	-0.00498***	-0.00449***	-0.00432***
	(0.00014)	(0.00014)	(0.00014)	(0.00015)	(0.00015)
constant	-4.324**	-4.455**	-4.915***	-5.351***	-5.352***
	(1.47)	(1.47)	(1.48)	(1.48)	(1.48)
FHG Dummies	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes
sector dummies	Yes	Yes	Yes	Yes	Yes
country dummies	Yes	Yes	Yes	Yes	Yes
N	547365	547365	547365	547365	547365
Pseudo R^2	0.031	0.039	0.044	0.056	0.057
AIC	696558	690492	687182	678551	677585
Log-likelihood	-348167	-345133	-343478	-339163	-338680
Precision (%)	65.12	66.07	66.41	67.30	67.38
Recall (%)	93.36	91.83	91.05	89.34	89.28
Specificity	13.13	18.07	19.98	24.57	24.91

The results from this check reveal that the statistics of the EPO patents markedly improve, since recall rises from less than 1% to 90%, at the cost of a relatively smaller amount of specificity. Nevertheless, the results are still more skewed for the EPO analysis than for its US counterpart, where recall and specificity are about equal.

We conclude that the low fit statistics for the EPO analysis are likely due to the much larger time window in which this analysis takes place, considering that later applicants' decisions can less reliably be explained by patent citation indicators. Notwithstanding, there remains a difference in the goodness of fit between the EPO and the USPTO analysis, which indicates that patent citations may just simply be better at explaining renewal for USPTO patent applications.

Appendix B: Employing a J-test to find an optimal patent citation indicator

In section 3 we analyzed the power of different citation indicators to explain patent value. Our analysis lead to two rankings of citation indicators, based on EPO and USPTO renewal. It is however unclear if all citation indicators explain the same processes that relate to patent value, while having a different efficiency. Alternatively, different citation indicators may describe different processes that relate to patent value. If the latter is true, then it may be worth examining models containing multiple citation indicators.

We can also test our proposition more formally by using a J-test which distinguishes between non-nested models (Davidson and Mackinnon, 1981). Unfortunately, the J-test rapidly becomes computationally complex if all observed models are non-linear (Mackinnon et al., 1983). Therefore, we use a linear representation of patent renewal in which we simply estimate the number of years a patent is renewed ($t_{renewal}$) as a measure of value (as was shown in Bakker, 2017). Patents cannot be renewed beyond 20 years, even if the applicant is willing to pay any fee for the renewal. Accordingly, a censored (i.e. Tobit) regression is used. In these regressions we still apply the same controls as we used in the Cox survival regressions. The sample of patents needs to be reduced in the same way as in the logistic regression, because EPO patents after 1993 did not have a chance to be fully renewed by the end of our observation period. The results for the Tobit analyses are shown in table 19 for EPO patents, and table 20 for USPTO patents, both of which reveal the same ranking as the logistic and Cox survival analyses.

Table 32: Estimates of the number of years an EPO patent is renewed for different citation indicators and using a Tobit regression. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) $t_{renewal}$ EPO	(2) $t_{renewal}$ EPO	(3) $t_{renewal}$ EPO	(4) $t_{renewal}$ EPO
Ln(1+EPO count)	1.432***			
	(0.016)			
Ln(1+USPTO count)		1.225***		
		(0.012)		
Ln(1+DOCDB count)			1.792***	
			(0.014)	
Ln(1+INPADOC count)				1.819***
				(0.013)
Ln(Applt. size)	0.132***	0.145***	0.130***	0.128***
	(0.0067)	(0.0067)	(0.0066)	(0.0066)
Co-patented	0.128	0.128	0.0941	0.101
	(0.090)	(0.089)	(0.088)	(0.088)
Applicant experience	-0.0164***	-0.0150***	-0.0129***	-0.0121***
	(0.00063)	(0.00063)	(0.00062)	(0.00062)
Constant	-57.51	-73.72	-70.11	-78.34
	(67.6)	(67.3)	(66.5)	(66.4)
σ	6.142***	6.119***	6.040***	6.030***
	(0.0095)	(0.0094)	(0.0093)	(0.0093)
IPC3 Dummies	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes
country dummies	Yes	Yes	Yes	Yes
N	286805	286805	286805	286805
Pseudo R^2	0.012	0.014	0.018	0.019
AIC	1617367	1615239	1607935	1606933
Log-likelihood	-808489	-807425	-803773	-803273

Table 33: Estimates of the number of years an USPTO patent is renewed for different citation indicators and using a Tobit regression. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) $t_{renewal}$ USPTO	(2) $t_{renewal}$ USPTO	(3) $t_{renewal}$ USPTO	(4) $t_{renewal}$ USPTO
Ln(1+USPTO count)	2.169***			
	(0.015)			
Ln(1+EPO count)		1.623***		
		(0.020)		
Ln(1+DOCDB count)			2.132***	
			(0.016)	
Ln(1+INPADOC count)				2.004***
				(0.015)
Ln(Applt. size)	0.192***	0.195***	0.179***	0.181***
	(0.0078)	(0.0079)	(0.0078)	(0.0078)
Co-patented	-0.157	-0.122	-0.250*	-0.264*
	(0.10)	(0.11)	(0.10)	(0.10)
Applicant experience	-0.0204***	-0.0238***	-0.0184***	-0.0179***
	(0.00069)	(0.00070)	(0.00070)	(0.00070)
Constant	-156.0	-126.8	-142.5	-150.5
	(104.7)	(106.5)	(105.1)	(105.3)
σ	9.516***	9.683***	9.556***	9.567***
	(0.014)	(0.015)	(0.014)	(0.014)
IPC3 Dummies	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes
country dummies	Yes	Yes	Yes	Yes
N	571816	571816	571816	571816
Pseudo R^2	0.029	0.023	0.028	0.027
AIC	2522713	2538693	2526575	2527657
Log-likelihood	-1261152	-1269142	-1263083	-1263624

We conducted the J-test in the same way as Smith and Maddala (1983). This entails that for each indicator i, the following model was set up: as a null hypothesis H0, the model with citation indicator i, shown in equation 1, is assumed to be the only correct model. As noted before there is censoring present, leading to the renewal time being limited to 20 years, even if the expression on the Right-Hand Side (R.H.S) is larger than 20.

$$\text{H0:} \quad t_{renewal} = \begin{cases} \beta_{cit,i}\ln(i+1) + \vec{\beta}_{controls,i}\vec{x}_{controls} + \varepsilon_i & R.H.S < 20 \\ t_{renewal} = 20 & R.H.S > 20 \end{cases} \tag{1}$$

The alternative hypothesis H1 states that some other model using a different citation indicator $j \neq i$ is correct instead. It is important to note that the J-Test assumes that one, and only one, model is correct. Therefore, the null hypothesis and the alternative hypothesis are exclusive.

$$\text{H1:} \quad t_{renewal} = \begin{cases} \gamma_{cit,j}\ln(j+1) + \vec{\gamma}_{controls,j}\vec{x}_{controls} + \varepsilon_j & R.H.S < 20 \\ t_{renewal} = 20 & R.H.S > 20 \end{cases} \tag{2}$$

To test whether the null hypothesis should be rejected, we construct an artificial regression that combines the models presented in H0 and H1, which will be

denoted by H_c. There is not one but several alternative models H1, thus requiring a sum of the models of the other indicators to create the combined model. This leads to the following combined regression to test the null hypothesis for indicator i, with weights λ_j belonging to each alternative model, as shown below.

$$H_c: \quad t_{renewal} = \left(1 - \sum_{j \neq i} \lambda_j\right) \left(\beta_{cit,i} \ln(i+1) + \vec{\beta}_{controls,i} \vec{x}_{controls}\right) + \sum_{j \neq i} \lambda_j \left(\hat{\gamma}_{cit,j} \ln(j+1) + \vec{\hat{\gamma}}_{controls,j} \vec{x}_{controls}\right) + \varepsilon_c \quad R.H.S < 20$$

$$t_{renewal} = 20 \quad R.H.S > 20$$

In this combined model, all coefficients belonging to the models, tested under the alternative hypotheses (i.e. $\hat{\gamma}_{cit,j}$ and $\vec{\hat{\gamma}}_{controls,j}$), are not estimated directly, but use fitted values obtained from estimating the model under H1 for each citation indicator j.

The J-test can be performed using this model by estimating the combined probability of the weights all being zero, i.e. $\lambda_j = 0$. Alternatively, we can perform a LR-test between the combined model and the model that is assumed correct under H0. In this subsection we perform both tests.

The J-test should only be used to test one model, thus in our case testing one indicator at a time .As such, we repeated the above procedure in which one indicator is set as i and the other indicators are set as j. The results of this exercise are shown in table 21 for EPO data, and in table 22 for USPTO data.

Table 34: Tobit regressions of combined models that estimate the number of years an EPO patent is renewed. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) $t_{renewal}$ EPO	(2) $t_{renewal}$ EPO	(3) $t_{renewal}$ EPO	(4) $t_{renewal}$ EPO
Ln(1+EPO count)	0.426***			
	(0.019)			
Ln(1+USPTO count)		-0.518***		
		(0.025)		
Ln(1+DOCDB count)			0.809***	
			(0.054)	
Ln(1+INPADOC count)				1.353***
				(0.044)
$\lambda_{EPO\ count}$		0.298***	0.298***	0.298***
		(0.013)	(0.013)	(0.013)
$\lambda_{USPTO\ count}$	-0.423***		-0.423***	-0.423***
	(0.020)		(0.020)	(0.020)
$\lambda_{DOCDB\ count}$	0.451***	0.451***		0.451***
	(0.030)	(0.030)		(0.030)
$\lambda_{INPADOC\ count}$	0.744***	0.744***	0.744***	
	(0.024)	(0.024)	(0.024)	
Ln(Applt. size)	0.0277***	-0.0730***	0.0470***	0.0840***
	(0.0068)	(0.0070)	(0.0083)	(0.0076)
Co-patented	0.0323	-0.0600	0.0368	0.0692
	(0.088)	(0.088)	(0.088)	(0.088)
Applicant experience	-0.00370***	0.00756***	-0.00463***	-0.00779***
	(0.00064)	(0.00067)	(0.00080)	(0.00072)
Constant	-12.85	35.46	-27.37	-54.00
	(66.2)	(66.2)	(66.3)	(66.2)
σ	6.016***	6.016***	6.016***	6.016***
	(0.0093)	(0.0093)	(0.0093)	(0.0093)
IPC3 Dummies	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes
country dummies	Yes	Yes	Yes	Yes
N	286805	286805	286805	286805
Pseudo R^2	0.020	0.020	0.020	0.020
AIC	160568	1605674	160568	160568
Log-likelihood	-802640	-802640	-802640	-802640
Test statistics				
$F(\forall\lambda = 0)$	3924***	3193***	755***	422***
LR compared to H0	11699***	9571***	2267***	1266***

Table 35: Tobit regressions of combined models that estimate the number of years an EPO patent is renewed. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) $t_{renewal}$ USPTO	(2) $t_{renewal}$ USPTO	(3) $t_{renewal}$ USPTO	(4) $t_{renewal}$ USPTO
Ln(1+USPTO count)	1.798***			
	(0.025)			
Ln(1+EPO count)		0.734***		
		(0.022)		
Ln(1+DOCDB count)			-0.391***	
			(0.055)	
Ln(1+INPADOC count)				0.704***
				(0.046)
$\lambda_{USPTO\ count}$		0.829***	0.829***	0.829***
		(0.012)	(0.012)	(0.012)
$\lambda_{EPO\ count}$	0.452***		0.452***	0.452***
	(0.014)		(0.014)	(0.014)
$\lambda_{DOCDB\ count}$	-0.183***	-0.183***		-0.183***
	(0.026)	(0.026)		(0.026)
$\lambda_{INPADOC\ count}$	0.351***	0.351***	0.351***	
	(0.023)	(0.023)	(0.023)	
Ln(Applt. size)	0.0533***	-0.0177*	-0.139***	-0.0420***
	(0.0085)	(0.0080)	(0.0098)	(0.0091)
Co-patented	-0.0432	0.0318	0.133	-0.00563
	(0.10)	(0.10)	(0.10)	(0.10)
Applicant experience	-0.00582***	0.000304	0.0144***	0.00477***
	(0.00078)	(0.00072)	(0.00092)	(0.00084)
Constant	-70.74	1.229	84.67	5.653
	(104.5)	(104.5)	(104.6)	(104.5)
σ	9.496***	9.496***	9.496***	9.496***
	(0.014)	(0.014)	(0.014)	(0.014)
IPC3 Dummies	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes
country dummies	Yes	Yes	Yes	Yes
N	571816	571816	571816	571816
Pseudo R^2	0.030	0.030	0.030	0.030
AIC	2520811	2520811	2520811	2520811
Log-likelihood	-1260198	-1260198	-1260198	-1260198
Test statistics				
$F(\forall\lambda = 0)$	634***	5895***	1926***	2286***
LR compared to H0	1908***	17888***	5770***	6852***

The J-test indicates that in all cases the null model needs to be rejected. It therefore shows that there is no model of patent renewal, which is absolutely better than the other models, given that in each combined regression the other models still explain a significant part of the variance in the patent renewal indicator. Consequently, we conclude that different citations capture different processes that correlate with patent value.

The J-test also allows for an additional verification of the ranking of the patent indicators. The combined model, in which the test statistic for the weights of the other models is lowest, likely contains the best null model. In our case this means that the model with the lowest F-statistic estimates directly the citation indicator that best explains patent renewal. In the case of EPO patents this entails the model containing the INPADOC count indicator, while in the USPTO model this entails the model containing the USPTO count indicator. Thereby confirming the results of the various horse-race regression analyses shown in the main body of the paper.

Appendix C: Does a 'home-bias' exist in patent citations?

Criscuolo (2006) found that patent counts of applicants could be skewed, when they are only evaluated in one patent system. It is possible that not only the patent counts have such a home bias, but that citation counts are also similarly skewed by applicant origin. A citation home bias is likely introduced because of two reasons: first, foreign applicants may simply write their patent documents differently, as well as follow a patent strategy that is more related to their home country; and second, foreign applicants may simply have fewer patents, leading to fewer self-citations. Furthermore, if the applicant is part of an industry which is clustered outside the country that the patent office resides in there may also be less citations from domestic applicants.

To investigate such a possible bias we will use the same set of patents, as used in the main body of the paper. We will denote applicants of patents by the country attributed to them in the patent application. The applicants are categorized in three groups: US applicants; European applicants, which belong to countries that subscribe to the EPO; and other applicants, which belong to any other country.

In these analyses we face the challenge of patents with multiple applicants, albeit that this only covers about 5% of our sample. To deal with this issue we assume the following: if the patent is applied for by a domestic and a foreign applicant, then the native applicant deals with the application process. Therefore, whenever a native applicant is present we ascribe the patent to that applicant. When the two non-domestic categories are both present we divide the patent between them using a partial count method. In tables 23 and 24 we give statistics for the EPO and USPTO patents in our sample, per the country of their applicant.

Table 36: Statistics of applicant origin for the number of USPTO patents in our sample. Any patent with an US applicant is solely subscribed to the US, while European (i.e. applicant belonging to a country which subscribes to the EPO), and other applicants, have a full distinct count.

Applicant origin	Number of USPTO patents	% of total USPTO patents	Average USPTO citations	Average DOCDB citations	Average INPADOC citations
US	182833	31.97	23.08	33.17	44.53
Europe	238196	41.66	9.84	15.16	15.23
Other	151930	26.57	12.49	18.63	21.03

Table 37: Statistics of applicant origin for the number of EPO patents in our sample. Any patent with an European, i.e. applicant belonging to a country which subscribes to the EPO, applicant is solely subscribed to Europe, while US and other applicants have a full distinct count.

Applicant origin	Number of EPO patents	% of total EPO patents	Average EPO citations	Average DOCDB citations	Average INPADOC citations
Europe	244,891	44.74	1.39	14.50	14.56
US	155,379	28.39	1.40	29.35	38.84
Other	147,896	27.02	2.26	17.67	19.89

From these tables we can already determine that, for EPO patents, there does not appear to be much of a difference between EPO and USPTO applicants regarding the EPO count. Foreign applicants however, receive more citations. This may indicate that the EPO system is less biased, as was already claimed in Criscuolo et al. (2005). The USPTO system does appear biased, with patents from US applicants cited at almost double the rate at which patents from other applicants are cited.

The family indicators also appear to be very biased towards US applicants. Patents with European applicants have low citation scores, followed relatively close by those of other applicants, while patents from US applicants are cited much more. This is likely due to the high correlation between the USPTO count and the DOCDB count indicators, which makes them similar enough that we can expect similar behavior. Nonetheless, this is only a simple tabulation, it is very well possible that (a part of) the bias is explained by quality and other differences between the patents. Therefore, we will conduct an analysis in which we control for this heterogeneity.

It will be necessary to control for the quality of a patents, when estimating a possible home bias. For instance, it is possible that domestic and foreign applicants write different patents, and produce patents of a different quality. For example, a firm willing to protect its invention outside of its home market may very well select only its best inventions to do so. To prevent such a selection effect from affecting our analyses we will include a set of controls that correct for patent quality, based on Squicciarini (2013), and which is also used in the third chaper of this PHD. Table 25 lists the controls and their descriptive statistics.

Table 38: Description and descriptive statistics of patent quality indicators used in this section. Statistics are given for USPTO patents and will vary only slightly for EPO patents.

indicator	Description	N	mean	Std. dev	min	max
Nr. Countries	Number of distinct patent offices in which the DOCDB family of the patent has at least 1 application present .	571816	7.08	4.12	2	51
Triadic	Dummy to indicate if the DOCDB patent family of the patent also contains a Japanese patent	571816	0.71	0.45	0	1
Backward citations	The number of patents cited by the focal patent	571816	11.88	12.67	0	198
Originality	The score of trajtenberg originality indicator on the IPC6 level	571816	0.51	0.32	0	1
Number of claims	The number of claims in the patent	571816	14.57	12.37	0	596
No claims registered	Dummy to indicate patents where the number of claims is unknown. For these patents the number of claims is set to 0	571816	0.0004	0.02	0	1
Grant lag	The number of days between the filing of the application document and the grant date of the patent	571816	798.65	430.99	0	9827
Number of IPC classes	Number of IPC technology classes that are assigned to the patent	571816	5.11	5.08	1	166
Number of distinct IPC3 classes	Number of distinct IPC3 classes that the patent is assigned to	571816	1.74	0.93	1	16
Shane radicalness	The score on the Shane radicalness at the IPC6 level	571816	10.05	11.31	0	198

We will evaluate the rate at which patents receive citations using a Poisson regression. Here, the number of citations the patent receives in its own office, will be the dependent variable. In both analyses we set applicants that do not originate from the US or Europe as the reference category. Because it is possible that the quality of a patent differs between applicants from different origins we also ran regressions where we control for the quality of the patent, by using the aforementioned control variables. The results of these Poisson regressions are listed in tables 26 and 27.

Table 39: Poisson regressions for granted EPO patent applications to explain the number of times they are cited based on the nationality of the applicant of the patent. Robust standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) EPO Count	(2) EPO Count	(3) EPO Count	(4) EPO Count
European applicant	-0.110*** (0.0064)	-0.311*** (0.0081)	-0.0641*** (0.0067)	-0.288*** (0.0081)
US applicant		-0.354*** (0.0083)		-0.383*** (0.0085)
Ln(1+firm size)	0.0878*** (0.0014)	0.0574*** (0.0015)	0.0838*** (0.0014)	0.0516*** (0.0015)
Co-patented	0.0422 (0.027)	0.0108 (0.027)	0.0179 (0.027)	-0.0136 (0.027)
Firm experience	-0.00407*** (0.00013)	-0.00155*** (0.00015)	-0.00351*** (0.00013)	-0.000894*** (0.00015)
Constant	-11.01*** (0.37)	-11.12*** (0.37)	-13.82*** (0.40)	-14.12*** (0.40)
Quality controls	No	No	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
IPC3 controls	Yes	Yes	Yes	Yes
Applicant controls	Yes	Yes	Yes	Yes
N	547365	547365	547365	547365
Pseudo R^2	0.071	0.076	0.093	0.098
AIC	2522735	2510540	2464175	2450711
Log-likelihood	-1261225	-1255126	-1231937	-1225200

Table 40: Poisson regression for granted USPTO patent applications to explain the number of times they are cited based on the nationality of the applicant of the patent.
Robust Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) USPTO count	(2) USPTO count	(3) USPTO count	(4) USPTO count
US applicant	0.677***	0.599***	0.571***	0.496***
	(0.0048)	(0.0066)	(0.0054)	(0.0071)
European applicant		-0.123***		-0.118***
		(0.0063)		(0.0064)
Ln(1+firm size)	0.0184***	0.00976***	0.0215***	0.0134***
	(0.0012)	(0.0013)	(0.0012)	(0.0013)
Co-patented	-0.0552*	-0.0592**	-0.0556*	-0.0594**
	(0.023)	(0.023)	(0.022)	(0.022)
Firm experience	-0.00311***	-0.00252***	-0.00282***	-0.00227***
	(0.00011)	(0.00012)	(0.00011)	(0.00012)
Constant	7.670***	7.479***	6.895***	6.730***
	(0.24)	(0.24)	(0.27)	(0.27)
Quality controls	No	No	Yes	Yes
year dummies	Yes	Yes	Yes	Yes
IPC3 dummies	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes
N	571816	571816	571816	571816
Pseudo R^2	0.207	0.208	0.243	0.244
AIC	11489829	11477769	10961940	10951294
Log-likelihood	-5744765	-5738733	-5480811	-5475486

From the results presented in tables 26 and 27 it indeed appears that a home bias exists: the origin of the applicant is a significant indicator of the number of patent citations, and this differs per office. USPTO patents from US applicants are significantly more cited than USPTO patents from EPO or other applicants, even after correcting for patent quality. At the EPO applicants from non-EU and non-US countries perform better, while patents from EPO applicants still get cited more than those from USPTO applicants. Correcting for patent quality appears to increase this difference even more.

The next question is whether family based indicators are also biased with respect to applicant origin. The descriptive statistics indicate that they may be biased in favor of applicants from the US. To analyze this question we again ran Poisson regressions but now with DOCDB and INPADOC patent counts as dependent variables (see table 28). Here we only present the results for the USPTO patents since the results for the EPO patents in our sample are very similar, due to the construction of our sample.

Table 41: Poisson regressions to determine a possible bias towards different applicants with either the DOCDB count or the INPADOC count as a dependent variable. The reference category is composed of patents from non-US and non-European origin. Robust standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) DOCDB count	(2) DOCDB count	(3) INPADOC count	(4) INPADOC count
US applicant	0.558***	0.373***	0.0167***	0.0242***
	(0.0070)	(0.0071)	(0.00197)	(0.00187)
EU applicant	-0.0945***	-0.111***	-0.0997*	-0.156***
	(0.0063)	(0.0061)	(0.046)	(0.0385)
Ln(1+firm size)	0.0151***	0.0176***	-0.00670***	-0.00562***
	(0.0015)	(0.0014)	(0.00021)	(0.000195)
Co-patented	-0.0631	-0.0896**	0.732***	0.508***
	(0.036)	(0.032)	(0.012)	(0.0117)
Firm experience	-0.00392***	-0.00326***	-0.183***	-0.160***
	(0.00012)	(0.00012)	(0.010)	(0.00996)
Constant	1.829***	-0.809***	5.694***	0.772*
	(0.21)	(0.24)	(0.34)	(0.394)
Patent quality	No	Yes	No	Yes
year dummies	Yes	Yes	Yes	Yes
IPC3 dummies	Yes	Yes	Yes	Yes
Sector dummies	Yes	Yes	Yes	Yes
N	571816	571816	571816	571816
Pseudo R^2	0.179	0.282	0.202	0.360
AIC	14918619.2	13044469.2	25699789.4	20620742.6
Log-likelihood	-7459167.6	-6522077.6	-12849745.7	-10310212.3

This analysis shows that family based indicators in themselves are still suffering from a bias towards US applicants. This analysis therefore confirms the previous indication that family indicators are as biased as USPTO citation indicators.

In conclusion, the least biased indicator found is the EPO count indicator. The USPTO count indicator, as well as the patent family based indicators, are biased towards applicants from the US. It is therefore problematic to create cross country comparisons using any of these three indicators. Moreover, it is likely that this bias will translate in reduced performance of the indicators to explain patent value. In the main body of this paper, we will also evaluate the efficiency of patent citation indicators to explain patent value depending on the origin of the applicants of the patents present in the sample.

Appendix D: Descriptive statistics and relevant multivariate analyses of exclusive citation indicators

In this appendix we discuss, in more depth, the descriptive statistics of the exclusive citation indicators that are described in the fourth section. We start in table 29 where we provide an overview of all patent citations to our sample, separated by patent office of origin.

Table 42: Patent citations to our sample denoted by the office of the citing patent. For patent offices, of patents not in the sample, offices are denoted by their country of residence if applicable. Percentages in bold add to 100%.

Office of citation	Patent citations given	Share of total
USPTO	*2765226*	*66.81%*
EPO	*551267*	*13.32%*
PCT	*540822*	*13.07%*
Europe	*177228*	*4.28%*
Germany	73127	1.77%
Great Britain	43699	1.06%
France	40955	0.99%
Spain	5976	0.14%
Netherlands	3720	0.09%
Italy	3611	0.09%
Austria	2321	0.06%
Belgium	1255	0.03%
Czechia	1238	0.03%
Switzerland	418	0.01%
Bulgaria	270	0.01%
Turkey	267	0.01%
Greece	211	0.01%
Luxemburg	131	0.00%
Denmark	22	0.00%
Norway	5	0.00%
Finland	2	0.00%
Other	*104230*	*2.52%*
Australia	60541	1.46%
Korea	20021	0.48%
Japan	17224	0.42%
Singapore	3074	0.07%
EAPO (Eurasian patents)	2628	0.06%
ARIPO (African Patents)	329	0.01%
Malaysia	300	0.01%
Russia	113	0.00%
Total	4138773	100%

It is to be noted that these patent statistics are not directly a projection of the patent office's themselves, but rather of the data that is available in the EPO PATSTAT database of their documents. This is especially true for the 'other citations' category as it holds only patents of a few offices. Table 29 indicates that the number of citations from each of these offices is rather low, which indicates that the coverage of some offices may not be complete.

Bakker et al. (2016) found that citation indicators could differ drastically, and very often had a low correlation. Therefore, we compute the correlations between the different partial citation indicators. Correlations of citation indicators, aggregated at the patent level, can be found in tables 30 and 31.

Table 43: Correlations between different patent citations to EPO patents and their family members from different sources. All correlations are significant at the $p<0.01$ level. N=547,365.

Variable	Nr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Patent-EPO	1	1														
Patent-USPTO	2	.6	1													
Patent-PCT	3	.8	.6	1												
Patent-Europe	4	.4	.0	.4	1											
Patent-Other	5	.4	.4	.4	.2	1										
DOCDB-EPO	6	.1	.1	.1	.0	.0	1									
DOCDB-USPTO	7	.2	.2	.2	.1	.1	.6	1								
DOCDB –PCT	8	.1	.1	.1	.0	.0	.9	.6	1							
DOCDB –Europe	9	.0	.0	.0	.1	.0	.4	.3	.3	1						
DOCDB –Other	10	.2	.2	.2	.1	.1	.4	.4	.4	.2	1					
INPADOC-EPO	11	.0	.0	.0	.0	.0	.2	.2	.2	.1	.1	1				
INPADOC –USPTO	12	.0	.0	.0	.0	.0	.2	.3	.2	.1	.1	.8	1			
INPADOC –PCT	13	.0	.0	.0	.0	.0	.2	.2	.3	.1	.1	1.0	.8	1		
INPADOC –Europe	14	.0	.0	.0	.0	.0	.2	.2	.2	.1	.1	.7	.8	.7	1	
INPADOC –Other	15	.1	.1	.1	.0	.0	.2	.2	.2	.0	.1	.8	.7	.7	.6	1

Table 44: Correlations between different patent citations to USPTO patents and their family members from different sources. All correlations are significant at the p<0.01 level. N=571,816.

Variable	Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Patent-EPO	1	1														
Patent-USPTO	2	.5	1													
Patent-PCT	3	.9	.5	1												
Patent-Europe	4	.2	.2	.2	1											
Patent-Other	5	.3	.3	.3	.2	1										
DOCDB-EPO	6	.3	.3	.3	.1	.1	1									
DOCDB-USPTO	7	.1	.2	.1	.0	.1	.6	1								
DOCDB –PCT	8	.3	.2	.3	.1	.1	.9	.6	1							
DOCDB –Europe	9	.1	.2	.1	.2	.1	.3	.2	.3	1						
DOCDB –Other	10	.2	.3	.2	.1	.2	.5	.4	.4	.2	1					
INPADOC-EPO	11	.1	.1	.1	.0	.0	.2	.2	.2	.0	.1	1				
INPADOC –USPTO	12	.1	.1	.1	.0	.0	.1	.2	.1	.0	.1	.8	1			
INPADOC –PCT	13	.1	.1	.1	.0	.0	.2	.2	.2	.0	.1	1.0	.8	1		
INPADOC –Europe	14	.1	.1	.1	.0	.0	.1	.2	.1	.1	.1	.7	.7	.6	1	
INPADOC –Other	15	.1	.1	.1	.0	.0	.1	.2	.1	.0	.1	.7	.8	.7	.7	1

These tables reveal that the correlations between citation sources are reasonably high, with correlations often being around 0.5. Yet, as already noted in Bakker et al. (2016), these correlations are low enough to show also that different citation sources still carry different information.

There is an extremely high correlation between EPO given citations and PCT given citations. This is caused by the fact that many EPO patents arrive through the PCT process and are examined by a patent examiner for the EPO. If such a patent is then followed up by an EPO application the backward citations found at in the previous examination are often re-used. Since this process occurs frequently many EPO and PCT documents share the same citations.

Incidentally, it is this process that causes problems with EPO backward citations in many databases, such as PATSTAT, as the backward citations of EPO patents that follow this PCT route are not well documented. In this paper we corrected this by adding the backward citations of the PCT documents to their affected EPO family members. However, because of this correction many backward citations from EPO and PCT documents are now equal, leading to the very high correlation between the indicators based on these documents. Because of this high correlation there is a concern for co-linearity between the EPO and PCT count if these are used as independent variables in the same analysis.

The correlation tables provide good indications for correlations between sets of citation indicators. However, to better understand the data we performed exploratory factor analyses (see tables 32 and 33 for EPO and USPTO patents respectively). These analyses reveal that citation indicators can often better be grouped by the documents (application, family) they refer to rather than the office they belong to, which appears as a second grouping mechanism. This finding confirms the approach followed in the previous section, where patent citations were mainly grouped by applications and different families.

Table 45: Exploratory factor analysis for patent citation indicators based on EPO patents. Varimax rotated solution where contributions above 0.4 are marked in bold.

	1(37%)	2(56%)	3(66%)	4(73%)	uniqueness
Patent-EPO	0.0567	**0.8979**	0.003	0.0601	0.187
Patent -USPTO	0.0908	**0.7701**	-0.0525	0.0846	0.3888
Patent -PCT	0.0942	**0.8624**	-0.0105	0.0159	0.247
Patent -Europe	-0.0655	**0.5571**	0.2547	0.0311	0.6195
Patent -Other	0.0077	**0.56**	-0.049	0.109	0.6721
DOCDB-EPO	**0.8946**	0.0511	0.1336	-0.0301	0.1784
DOCDB-USPTO	**0.675**	0.1533	0.109	0.2336	0.4544
DOCDB -PCT	**0.8988**	0.0569	0.0974	-0.0481	0.1771
DOCDB -Europe	0.2069	0.0123	**0.946**	0.0131	0.0619
DOCDB -Other	0.296	0.1468	0.0923	**0.807**	0.231
INPADOC -EPO	**0.8893**	-0.0126	0.1014	0.1782	0.167
INPADOC-USPTO	**0.7082**	0.0496	0.1022	0.3863	0.3363
INPADOC -PCT	**0.9044**	-0.0087	0.0725	0.1453	0.1557
INPADOC -Europe	0.3234	-0.0174	**0.8916**	0.1485	0.0781
INPADOC -Other	**0.4317**	0.0604	0.0969	**0.8512**	0.076

Table 46: Exploratory factor analysis for patent citation indicators based on USPTO patents. Varimax rotated solution where contributions above 0.4 are marked in bold.

	1(38%)	2(53%)	3(63%)	4(71%)	5(76%)	uniqueness
Patent-EPO	0.0875	**0.9125**	0.1342	0.0097	0.0088	0.1416
Patent -USPTO	0.0988	**0.6521**	0.0647	0.1454	0.2665	0.4687
Patent -PCT	0.0878	**0.9111**	0.1455	0.0002	0.0033	0.1409
Patent -Europe	-0.0562	**0.3982**	-0.1204	**0.3841**	0.0623	0.6723
Patent -Other	-0.0602	**0.4408**	-0.0617	0.106	**0.3813**	0.6416
DOCDB-EPO	0.2864	0.1421	**0.8726**	0.1819	0.1733	0.0733
DOCDB-USPTO	**0.4105**	0.0273	**0.5674**	0.1029	0.1921	0.4613
DOCDB -PCT	0.2988	0.1446	**0.8828**	0.1297	0.1399	0.074
DOCDB -Europe	0.0026	0.024	0.2128	**0.9407**	0.0647	0.0651
DOCDB -Other	0.1014	0.051	**0.3069**	0.0823	**0.8863**	0.1007
INPADOC -EPO	**0.8395**	0.1122	**0.3719**	0.1539	0.1069	0.1092
INPADOC-USPTO	**0.8965**	0.0288	0.1382	0.1195	0.1549	0.1381
INPADOC -PCT	**0.8274**	0.1174	**0.4047**	0.1149	0.0803	0.1182
INPADOC -Europe	**0.3594**	0.0284	0.0797	**0.8883**	0.0834	0.0676
INPADOC -Other	**0.5985**	0.0444	0.0573	0.1072	**0.7246**	0.0999

Appendix E: The effects of intra family competition

The analysis of the fourth section showed that citations to patent family members are positively related to the chance a patent is abandonned. A possible explanation for this result is based on the existence of competition between patents that belong to the same patent family.

It is likely that patents from the same patent family protect the same invention- for example, by protecting several components and/or applications of one invention. Patents belonging to the same DOCDB patent family tend to protect the same invention in multiple jurisdictions, while patents form the same INPADOC patent family may protect different components of the same invention.

When innovators file patents to protect their inventions, they may be inclined to try and protect as much of their efforts as possible, since they are uncertain of the reaction of potential imitators. After some time has passed, and the costs for protection have increased due to increased maintenance fees, innovators may decide to abandon some of their patents because they have become redundant in protection. This then lead to competition between patents belonging to the same patent family.

Patents that have a larger family will be more likely to have a higher family citation score because there are more patents that can be cited. Additionally, if there is a patent in a family that is cited more often it may have a higher relative performance. In this case the family members with lower citation scores should be less likely to be maintained. Therefore, competition may lead to the abandonnement of patents, whose family members have more citations.

To investigate the veracity of this competition model we repeated the same analysis as performed in table 13. Here, we introduce two new variables: the number of other granted applications, in the same family and in the same office, as well as their combined citation count. Furthermore, the family citation count in the same office will be changed to only include patent citations from outside the patent office. This allows us to answer the research question by testing the hypothesis that the coefficient of the count of granted patents in the same family and office, is positively related to the chance of abandonment. A similar hypothesis can bet tested regarding the number of citations received by these patent family members.

For our sample, there are 43,121 USPTO patents with granted USPTO family members in their DOCDB family, and 69,454 USPTO patents with granted USPTO family members in their INPADOC patent family. We estimate the effects of the number of other USPTO granted patents in the DOCDB or INPADOC family on the renewal time of the USPTO patent. Additionally, we also include the number of citations to these patents, as competition is also

dependent on the value of the other patents. These citations have then been subtracted from the DOCDB-USPTO indicator to avoid double counting. The results of this exercise are listed in table 34.

Table 47: Cox survival regression to test the effects of within family competition for granted USPTO patents. Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

	(1) USPTO renewal time	(2) USPTO renewal time
Other patents in same family	-0.000452 (0.0014)	0.000865*** (0.00019)
Citation counts to these other patents	0.000358*** (0.00010)	-0.000239*** (0.000069)
Patent-USPTO	-0.0153*** (0.00016)	-0.0153*** (0.000158)
DOCDB-USPTO	0.000827*** (0.00023)	
INPADOC-USPTO		-0.000282*** (0.000075)
Ln(Applt. size)	-0.0239*** (0.0010)	-0.0239*** (0.0010)
Co-patented	0.0228 (0.013)	0.0229 (0.013)
Applicant experience	0.00238*** (0.000092)	0.00238*** (0.000092)
Year dummies	Yes	Yes
IPC3 dummies	Yes	Yes
Sector dummies	Yes	Yes
country dummies	Yes	Yes
N	571816	571816
Pseudo R^2	0.007	0.007
AIC	7506694	7506651
Log-likelihood	-3753141	-3753119

We find support for the first hypothesis: the presence of patent family members in the same office decreases the chances of patent renewal. The associated coefficient is very significant for INPADOC patent family members, and insignificant for DOCDB patent family members.

The second hypothesis, which states that citations to these patent family members also decreases the chances a patent is renewed, is only confirmed for DOCDB patent family members. More citations to INPADOC patent family members are associated with an increased chance of renewal. This may indicate that patent citations are mainly a good method to gauge competition between patents that have the same technical content.

In conclusion, we find that patents may have a reduced chance of being renewed if there are patent family members present in the same office. This finding reveals that patent owners consider the place of a patent in their portfolio, when they make decisions regarding patent maintenance. This is not commonly addressed in the patent literature and could provide an interesting avenue to better understand how portfolios are managed.

References

Bakker, J. The log-linear relation between patent citations and patent value. *Scientometrics*, *110(2), 879-892*.

Bakker, J., Verhoeven, D., Zhang, L., and Van Looy, B. (2016). Patent citation indicators: One size fits all? *Scientometrics*, *106*(1), 187-211.

Criscuolo, P. (2006). The 'home advantage' effect and patent families. A comparison of OECD triadic patents, the USPTO and the EPO. *Scientometrics*, *66*(1), 23-41.

Criscuolo, P., Narula, R., and Verspagen, B. (2005). Role of home and host country innovation systems in R&D internationalisation: a patent citation analysis. *Economics of Innovation and New Technology*, *14*(5), 417-433.

Czarnitzki, D., Hussinger, K., and Schneider, C. (2011). Commercializing academic research: the quality of faculty patenting. *Industrial and Corporate Change*, *20*(5), 1403-1437.

Czarnitzki, D., Hussinger, K., and Schneider, C. (2011). "Wacky" patents meet economic indicators. *Economics Letters*, *113*(2), 131-134.

Davidson, R., and MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica: Journal of the Econometric Society*, 781-793.

Dahlin, K. B., and Behrens, D. M. (2005). When is an invention really radical? Defining and measuring technological radicalness. *Research Policy*, *34*(5), 717-737.

Gambardella, A., Harhoff, D., and Verspagen, B. (2008). The value of European patents. *European Management Review*, *5*(2), 69-84.

Gay, C., and Le Bas, C. (2005). Uses without too many abuses of patent citations or the simple economics of patent citations as a measure of value and flows of knowledge. *Economics of Innovation and New Technology*, *14*(5), 333-338.

Graham, S. J., and Harhoff, D. (2006). Can post-grant reviews improve patent system design? A twin study of US and European patents.

Hall, B. H. (2004). Exploring the patent explosion. *The Journal of Technology Transfer*, *30*(1-2), 35-48.

Hall, B. H., Jaffe, A. B., and Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools* (No. w8498). National Bureau of Economic Research.

Harhoff, D., Scherer, F. M., and Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research policy*, *32*(8), 1343-1363.

Hottenrott, H., Hall, B. H., and Czarnitzki, D. (2016). Patents as quality signals? The implications for financing constraints on R&D. *Economics of Innovation and New Technology*, *25*(3), 197-217.

MacKinnon, J. G., White, H., and Davidson, R. (1983). Tests for model specification in the presence of alternative hypotheses: Some further results. *Journal of Econometrics*, *21*(1), 53-70.

Magerman, T., Van Looy, B., and Song, X. (2006). *Data production methods for harmonized patent indicators: Patentee Name Harmonization*. EUROSTAT Working Paper and Studies, Luxembourg.

Magerman, T., Van Looy, B., and Debackere, K. (2011, January). In search of anticommons: patent-paper pairs in biotechnology. An analysis of citation flows. In *4th ZEW Conference on Economics of Innovation and Patenting*.

Maurseth, P. B. (2005). Lovely but dangerous: The impact of patent citations on patent renewal. *Economics of Innovation and New Technology*, *14*(5), 351-374.

Martínez, C. (2010). Insight into different types of patent families. OECD Science and Technology Working Paper 2010/2.

Michel, J., and Bettels, B. (2001). Patent citation analysis. A closer look at the basic input data from patent search reports. *Scientometrics*, *51*(1), 185-201.

Nakamura, H., Suzuki, S., Kajikawa, Y., and Osawa, M. (2015). The effect of patent family information in patent citation network analysis: a comparative case study in the drivetrain domain. *Scientometrics*, *104*(2), 437-452.

Neuhäusler, P., and Frietsch, R. (2012). Patent families as macro level patent value indicators: Applying weights to account for market differences. Scientometrics, 96(1), 1–23

Sapsalis, E., de la Potterie, B. V. P., and Navon, R. (2006). Academic versus industry patenting: An in-depth analysis of what determines patent value.*Research Policy*, *35*(10), 1631-1645.

Schmoch, U. (2008). Concept of a technology classification for country comparisons. Final report to the world intellectual property organisation (wipo), WIPO.

Smith, M. A., and Maddala, G. S. (1983). Multiple model testing for non-nested heteroskedastic censored regression models. *Journal of Econometrics*, *21*(1), 71-81.

Webb, C., Dernis, H., Harhoff, D., and Hoisl, K. (2005). Analyzing European and International Patent Citations: A Set of EPO Patent Database Building Block. STI Working Paper 2005/9, OECD.

General conclusion

This doctoral dissertation presented several ways to improve the patent citation indicator, and thereby improving the measurement of innovation. Chapters 1 and 4 emphasize the importance of choosing a data source from which to extract the indicator. Chapter 2 details the necessity of a new functional form to relate patent citations to patent value. And finally, chapter 3 shows a new interpretation of patent citations, and why their correlation with patent value emerges. The research shown in all chapters provides guidance for those interested in using patent citations to approximate patent value.

Researchers using patent citations should heed the results in this thesis, when creating patent indicators for their own purposes. As can be seen in chapter 1, not all citations are equal, and it is likely that different citations can produce markedly different results, even if they all correlate positively with patent value (see chapter 4). Furthermore, those wishing to compare the current or past innovative performances of economic actors (individuals, firms, nations, etc.) need to be aware that all citation indicators carry a bias towards applicants of certain countries (often the U.S.). Even citation indicators that account for global patent families are biased towards the U.S, albeit that this bias appears relatively minor compared to that of patent citations based on single USPTO documents.

This thesis could also be of use for managers of intellectual property considering the framework presented in chapter 3. This chapter shows, for the first time ever, that patent citations may serve as a useful indicator of possible interest in intellectual property. Likewise, backward patent citations may also serve as a warning for potential litigation, and perhaps the need to find licensing agreements. For researchers interested into IP management, the network of patent citations may reveal the market structure of intellectual property.

Nonetheless, for researchers willing to apply the results from this thesis, there is a major hurdle to be taken, which relates to the availability of patent data. Currently, there is only one patent database readily and freely available (i.e. the NBER patent database). This database only covers patents and citations from US patents. To create the family based indicators, referred to in this thesis, it is necessary to obtain a larger database, such as the PATSTAT database. The costs of the database itself are relatively minor, but the skill required to create correct indicators takes quite some time to acquire. Consequently, it is not trivial to expect every researcher to adopt family based citation indicators, even if they are proven to be superior on their dataset.

There are two practical solutions to this problem. First, correct citation indicators could be made available on a centralized and easily accessible platform. This should not impose enormous difficulties, seeing as patent offices, such as the EPO, already calculate the forward citations per patent in

their online database: the only extra step would be to compile a few more indicators and make them available in an easily accessible table. If this is not possible, at least a table detailing which patents belong to which families can be made available. The results of chapter 1 suggest that researchers can aggregate patent citations at the patent family level, even when using patent information from a single office, since the resulting indicator will be reasonably similar to a patent family indicator which uses citation data from multiple offices. Making a simple table of patent family membership available would alleviate the issue substantially.

The insights of this thesis could also affect the construction of other patent indicators. For instance, the Trajtenberg et al. (1997) generality indicator depends heavily on patent citations, and could be substantially altered if a family based citation indicator is used, instead of a patent citation indicator based on patents of a single office.

Additionally, while patent citations often refer to the number of times a patent is cited, it may also be interesting to look at the sources cited by the patent. The patents that are cited in a patent document tend to lend themselves to various indicators of the citing patent such as: the Trajtenberg et al. (2007) originality indicator; the backward citations indicator (i.e. the count of patents cited); and the Verhoeven et al. (2016) new origins indicator. There are even indicators based on cited non-patent-literature. For instance, Van Looy et al. (2007) have found that academic sources cited by patents could indicate that the patent relies on more scientific knowledge. It would be an interesting endeavor to see if grouping these backward citations at the patent family level yields new, or better, insights. Furthermore, backward citations could also be grouped by the type of patent they cite, whether it belongs to the same owner, or if it is filed in the same patent office. Given the results presented in this dissertation, this could provide valuable insights in the position of the patent.

Another future research direction would be to investigate the value of patent portfolios instead of single patent documents. Chapter 2 already introduces a simple portfolio indicator derived from the value of single patent documents, but this does not take all the characteristics of the portfolio into account. To arrive at a good portfolio indicator the results of this thesis should be combined with portfolio concepts, such as the diversity and strategic position of the portfolio. In addition, the overlap in the portfolio, which could be measured by the self-citations it contains, could prove an interesting indicator for the effective protection it lends against imitation of the technologies of the patent owner.

Another theme is present throughout this thesis, which is the concerns the proper validation of constructs and concepts when measuring innovation. Too often indicators are simply assumed to measure a construct purely based on a theoretical argument, but without any validation being provided. This thesis shows the possibility of a thorough validation as well as some of the surprising results that can be achieved by doing so. For instance, a priori, aggregating received citations on the level of the patent familym rather than the patent itself, could provide a better indication of the value of the patent, since more of the available information is used. However, the results presented in chapter 4 show that this is not always the case.

In conclusion, this thesis shows that it is helpful to view patent citations for what they are: legal instruments that provide information for the citing patent. For example, chapter 3 shows that this conceptualization leads to interesting observations regarding the distinct types of value that can be represented by distinct types of citations. The other chapters show that our current use of patent citation indicators can and should be, improved. We can only hope that other scholars apply the lessons provided and arrive at a better understanding of patent citations, doing so will further the understanding of innovation and intellectual property.

References

Trajtenberg, M., Henderson, R., and Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and new technology*, *5*(1), 19-50.

Van Looy, B., Magerman, T., and Debackere, K. (2007). Developing technology in the vicinity of science: An examination of the relationship between science intensity (of patents) and technological productivity within the field of biotechnology. *Scientometrics*, 70(2), 441-458.

Verhoeven, D., Bakker, J., and Veugelers, R. (2016). *Measuring technological novelty with patent-based indicators. Research Policy,* 45(3), 707-723.